Deaf students consistently score lower on standardized measures of reading comprehension than their hearing peers. Most of the studies that have been conducted to explain this phenomenon have focused on variables within the reader, and important differences have been found between deaf and hearing readers. More recently, in the face of increasingly high-stakes consequences, researchers are looking “outside” the reader, at the tests themselves, to determine whether there are fairness issues for special populations, such as deaf students. The study reported here, the first of its kind with deaf students, examines the North Carolina (NC) reading comprehension test. The study employs the same method used originally by NC to determine its appropriateness of the test for the general population of NC students. The experts in this article, like those in the original construction of the NC test, are familiar with the content of the reading curriculum in NC; however, the raters in this article bring a special perspective related to teaching and testing reading of students who are deaf. Findings from this study raise questions about the appropriateness of the NC reading test for deaf students. Implications for future research and instructional practice are discussed.
that the tests were not designed for special education students (Cawthon, 2004, 2006, 2007; Martin & Mounty, 2005).

A consistent finding of comparative studies of deaf and hearing age-mates on standardized reading comprehension test performance for the past 90 years is that deaf students’ scores are much lower than those of hearing peers (Allen, 1986; DiFrancesca, 1972; Furth, 1966; Fusfeld, 1955; Goetzinger & Rousey, 1959; Holt, 1993; Karchmer & Mitchell, 2003; Lane & Baker, 1974; Pintner & Patterson, 1916; Pugh, 1946; Traxler, 2000; Trybus & Karchmer, 1977; Wrightstone, Aronow & Moskowitz, 1963). Karchmer and Mitchell (2003) report a gap of 6 years for the average reading comprehension performance of 15-year-old deaf and hearing students on the ninth edition of the Stanford Achievement Test. This is essentially the same gap that existed in the mid-1960s (Traxler, 2000). As will be discussed later in this article, most of the research conducted to explain the achievement gap between deaf and hearing readers has focused on “reader” variables, including linguistic and cognitive variables. Findings from these studies consistently show that deaf students tend to score lower than their hearing peers on tests of phonological awareness, vocabulary, syntax, and figurative language, inferential abilities, and working memory (See Paul, 2003 for a review of this research). In addition, as will be discussed later in this article, differences in reading test performance between deaf and hearing peers have been linked to “test” or “task” variables, including the length and language of the question stem, the ability to reinspect the text while answering questions, and the nature of the task (e.g., production vs. recognition) (LaSasso, 1979; LaSasso & Davey, 1983).

Increased Emphasis on Standardized Reading Tests and the Concept of Test Fairness

Increasingly, states are constructing and administering standardized reading competency tests to all students in the state, including students who are deaf (Cawthon, 2004, 2006; Martin & Mounty, 2005). This is due, in large part, to accountability measures mandated by the No Child Left Behind Act of 2001 (2002) and the Individuals with Disabilities Education Act Amendments of 1997 (2001). The goal of state-mandated tests is to raise both the level of accountability among educators and the quality of education for students; however, these tests often have “high-stakes” consequences for students, especially when students need to pass a reading competency test in order to receive a high school diploma; therefore, it is incumbent upon test developers, including developers of statewide mandated reading competency tests, to address psychometric standards related to the validity, reliability, and fairness of their test for the populations who are required to take the test, including students with limited English proficiency (Abedi, 2004; Geisinger, 1998; NRC, 1999) and students who are deaf (Cawthon, 2004, 2006, 2007; Martin & Mounty, 2005).

A consensus view of the psychometric properties necessary for defensible high-stakes test uses is described in the 1999 “Standards for Educational and Psychological Testing.” These standards were developed by three professional organizations concerned with testing: the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. Test fairness, defined as a lack of bias for different populations who take the test (Berk, 1982; Kunnan, 2008; NRC, 1999; Scheuneman & Oakland, 1998), is one of the psychometric properties addressed in the NRC standards. During the past three decades, test fairness has received considerable attention in the courts, legislatures, schools, and employment settings (Pratt & Moreland, 1998). According to the NRC (1999), “When test use is inappropriate, especially in making high-stakes decisions about individuals, it can undermine the quality of education and equality of opportunity. This lends special urgency to the requirement that test use with high-stakes consequences for individual students be appropriate and fair” (NRC, 1999, p. 4).

A criterion for a test being “fair” or “unbiased” is that the meaning of the test scores does not differ across individuals, groups, or settings (NRC, 1999); that is, if one population, based on gender, race–ethnicity, or family income, for example, is favored over another on a test item, the item is not fair and the test is considered to be biased.
**Item analysis.** Test fairness, or bias, is typically investigated via analysis of test items, utilizing one or both of the following methods: (a) a panel of experts familiar with the content of the curriculum and characteristics of the students who will take the test judges individual passages and questions for test bias or (b) individual test items on the test are subjected to a statistical procedure called differential item functioning (DIF) (Holland & Thayer, 1988; Mantel & Haenszel, 1959), in which test items are compared for two groups (i.e., a reference group and a focal group), which have been matched on overall test performance (Holland & Thayer, 1988). When panels of experts are used to determine whether test bias exists for a group of test takers, aspects of a test that are typically evaluated to determine its fairness for the students typically include: (a) the test itself, including the language, content, and test format and (b) uses of test results (e.g., for purposes such as tracking, promoting students, or awarding a diploma) (NRC, 1999).

**Language.** Much of the discussion in the professional literature about language bias on standardized reading tests for different populations relates to second language learners, or English language learners, who are less than proficient with the English language (Abedi, 2004; Abedi & Dietel, 2004; Abedi, Leon, & Mirocha, 2001; Bailey, 2000; Butler and Stevens, 2001; Cunningham & Moore, 1993; Kopriva, 2000, Messick, 1989; Sandoval & Duran, 1998). Abedi et al. (2001) and Abedi and Dietel (2004) found that English language learners score considerably lower on standardized reading tests than their native speaking peers. Abedi (2004) found that results on content reading assessments typically correlate with students’ language proficiency level; that is, the lower the student’s language proficiency, the lower they will likely score on content assessments. Kopriva (2000) and Mohan (1982) note that English language learners may not be able to demonstrate the depth of their knowledge and comprehension under the restrictions of large-scale tests that have been designed for mainstream students who share a common culture related to test taking. Specifically, when second language learners are not familiar with the mainstream culture, they are more apt to interpret questions differently than native speakers who are familiar with the host culture (Mohan, 1982). Bailey (2000) notes that the language found in reading tests is often beyond the proficiency level of many of the English language learners. In an experimentally designed study, Cunningham and Moore (1993) found that when the language of a reading test was modified for English language learners, their performance increased; that is, when everyday language and vocabulary was used in written comprehension questions, rather than academic language and vocabulary, the students’ performance increased. Cunningham and Moore speculate that English language learners may become frustrated when, despite having mastered the content knowledge being assessed, they are stymied by the language requirements of the test (Cunningham & Moore, 1993).

In many regards, deaf readers are like readers who are English language learners. Numerous studies have compared vocabulary, figurative language, and syntax of deaf and hearing peers on tests of written language. Findings from these studies consistently reveal that deaf students lag behind their hearing age-mates in each of these linguistic dimensions. In addition, the length and linguistic complexity of reading test items have been shown to differentially influence deaf and hearing students’ performance on reading tasks. Trybus and Buchanan (1973) found a positive correlation between the difficulty of the question for deaf students and the number of words in the question stem (i.e., probe or question) as well as the length of the passage. Rudner (1978) analyzed the data used by Trybus and Buchanan (1973) and concluded that the following linguistic structures created more difficulty for deaf test takers than their hearing peers when matched for reading level: (a) conditionals (e.g., “if” clauses), (b) comparatives (e.g., “greater than, the most”), (c) negatives (e.g., “not, without, answer not given”), (d) inferentials (“should, could, because, since”), (e) low information pronouns (e.g., “it, something”), and (f) lengthy passages. DiFrancesca and Carey (1972) found that differential performance between deaf and hearing readers on multiple-choice reading tests was related to some distracters not being as effective for deaf readers. For example, if one of the distracters were phonically similar to the correct answer, a hearing reader might select it, whereas a deaf
reader might not. Taken together, findings from these studies indicate that it is important to examine any high-stakes test for possible language bias for deaf students.

Content of the test. The content of a standardized reading test can also influence a student’s performance and, therefore, its fairness. Content “bias” exists when the content of the test does not match the content of the curriculum (NRC, 1999). It also exists when the content (i.e., of reading comprehension passages) assumes general world knowledge on the part of the student (LaSasso, 1979). To the extent that the content of the test does not match the curriculum, a student’s performance may be affected (Benning, 2000). A form of content bias, referred to as “fund of information bias,” exists when the content of an achievement test reflects the test-taker’s exposure to and familiarity with background knowledge, or general world knowledge, such as that which is derived from the radio or overhearing others’ conversations, as opposed to what is learned in school (Brauer et al., 1998, p. 305).

Due to reduced sensory input, deaf children’s fund of world knowledge is typically less extensive than that of hearing peers (Oakhill & Cain, 2000). Accordingly, it is imperative that test constructors assess their tests for content bias if those tests are to be used with deaf students, especially if those tests have high-stakes consequences, such as high school graduation.

Format. The “format” of a reading test, including the nature of the “stimulus” (e.g., passage and question type) and “response mode” (i.e., recognition vs. production) can also impact on a test-taker’s performance and a test’s fairness. Most standardized reading tests use a multiple-choice format with either WH-question stems (e.g., “who, what, where, when, why, how”) or incomplete statement stems (e.g., “The animal that the family saw was a ___”). Students’ performance on multiple-choice tests is influenced by a number of variables, including their overall experience taking multiple-choice tests (Frisby, 1998).

Research comparing deaf and hearing test takers on format variables (Bornstein, 1971; Davey & LaSasso, 1984; Davey, LaSasso, & Macready, 1983; DiFrancesca & Carey, 1972; LaSasso, 1979; McKee & Bondi-Wolcott, 1982; Osguthorpe, Long, & Ellsworth, 1977; Rudner, 1978; Trybus & Karchmer, 1977) reveals that a number of test, or task, variables differentially effect deaf and hearing peers’ reading comprehension test performance. These variables include: (a) the ability to reinspect, or look back at, the text while answering questions, (b) the production versus recognition response mode of the task, and (c) the placement of the correct answer in the list of options on multiple-choice tests (See King & Quigley, 1985; LaSasso, 1999; Paul, 1998 for a review).

One test variable that has been found to differentially impact on deaf and hearing test-taker’s performance is the ability to reinspect, or utilize, the text while responding to reading comprehension questions. Reinspection typically simplifies the task of answering questions because the reader does not need to rely as much on memory to answer a question about what was read and can reread the text as needed to answer the question (Garner & Reis, 1981); however, Davey et al. (1983) found that the opportunity to reinspect the text while answering multiple-choice questions benefited hearing participants more than deaf peers, as reflected in the lower scores of participants who are deaf than their hearing peers on the reinspection task.

A second test variable shown to differentially influence deaf and hearing test-takers’ performance on reading tests is the “production” versus “recognition” aspect of the task. A number of studies of deaf students’ test-taking abilities reveal that the production-recognition aspect of the reading task appears to influence the performance of deaf participants more than it does for hearing peers (Davey & LaSasso, 1984; Davey et al., 1983; LaSasso, 1979; McKee & Bondi-Wolcott, 1982; Osguthorpe et al., 1977). Findings from these studies, taken collectively, suggest that readers who are deaf, like readers who are hearing, perform better on reading comprehension tasks where they need to merely “select” a correct response from a number of options instead of “producing” responses; however, based on their reading test performance, readers who are deaf, do not appear to benefit as much as readers who are hearing from the opportunity to select rather than produce a response.
The position, or placement, of the correct answer among the foils, or distractors, on multiple-choice reading tests is a third test variable that has been shown to differentially affect deaf and hearing test-takers’ performance on tests. Trybus and Buchanan (1973) found that, unlike hearing participants in their study, deaf age-mates were more likely to get an item correct if the correct response were either the first or last in the list of options. Similar “position” effects were found in deaf college students by McKee and Bondi-Wolcott (1982) on an eight-option multiple-choice test. DiFrancesca and Carey (1972) also found that order of items affected deaf participants’ performance on reading tests administered to deaf and hearing students; that is, unlike their hearing peers, deaf participants answered more questions correctly at the beginning of the test than at the end.

In addition to the differential influence of language, content, and test format on deaf and hearing test-taker’s performance, students who are deaf have been found to use specific “compensatory” test-taking strategies when they are, presumably, not comprehending what they are reading (LaSasso, 1985; 1986; Webster, Wood, & Griffiths, 1981; Wolk & Schildroth, 1984; Wood, Griffiths, & Webster, 1981). Findings of Webster et al. (1981), Wolk and Schildroth (1984), and Wood et al. (1981) indicate that deaf participants, compared to hearing peers, tend to guess more often on multiple-choice tests. The same guessing phenomenon has been observed in students who are deaf on cloze tests where every fifth word of a passage has been deleted and replaced by a 10-letter underline. The reader’s task is on a cloze test to write the one word in each blank that the reader believes the author used (LaSasso, 1985, 1986). Webster et al. (1981) and Wood et al. (1981) observed a similar pattern on modified cloze tests where British deaf students were given a “cue sentence,” such as, “Pack the eggs in the box. Hens lay __” (Webster et al., 1981, p. 138). Webster et al. found that deaf participants were much more likely to fill in the blanks than hearing participants. Strategies included: (a) visual matching, (b) elimination of unlikely distractors, (c) word/idea association, and (d) selection of correct item based on position among choices on multiple-choice test. Guessing has also been found to be a factor in test scores of women and minority examinees, particularly Black examinees (See Scheuneman & Oakland, 1998 for a review).

Visual matching is one of the test-taking strategies that students who are deaf have been found to use extensively on short answer reading tests when reinspection of the passage is permitted (LaSasso, 1985, 1986). The strategy, presumably employed when a test taker is not able to answer a test question, involves responding by writing a verbatim word or series of words that occur within two lines above or below a word in the text that matches a word or words in the question (LaSasso, 1985). In a study of 50 participants who are deaf matched for reading level with 50 hearing participants, 76% of students who are deaf used the strategy at least 25% of the time, 33% used it at least 50% of the time, and 17% of the students used it at least 67% of the time. The average number of responses suggesting visual matching on the 24-item test used in that study was 4.2 for participants who are deaf and 0.1 for hearing participants matched for reading level with the deaf participants (LaSasso, 1985). Similar visual matching strategies have been observed with deaf readers in Greece (Savvides & LaSasso, 1988) and with deaf and hearing readers matched for both age and reading level (LaSasso, 1986). See Figure 1 for an example of visual matching.

In sum, due to the well-documented differential effect of language, content, and format variables for deaf and hearing peers on standardized reading tests in the United States, it is vital that constructors of reading competency tests establish the fairness of those tests for all groups taking the test, including students who are deaf. The same procedures that are typically used to establish test fairness with subgroups of test takers based on gender or race–ethnicity should...
be used to establish test fairness with the subgroup of test takers who are deaf or hard of hearing.

The North Carolina Reading Competency Test

*Initial construction of the North Carolina Reading Competency Test.* The North Carolina (NC) Reading Competency Test (Form M) consisted of 10 short passages and 68 questions. Of the 10 passages, four were “literary” (i.e., poetry, fiction, biographies, plays, essays), four were “content based” (i.e., science, social studies, art, health, and mathematics), and two were “consumer/human interest” (i.e., recipes, directions, forms, projects, brochures, and short informational pieces relevant to the students) (North Carolina Department of Public Instruction [NCDPI], 1996). Each multiple-choice test item contained a WH-question (i.e., began with who, what, when, where, why, or how) and four options consisting of the correct answer and three distracters or foils.

The NCDPI began test construction with the selection of 100 reading passages. Most passages used were excerpts from longer literary pieces; therefore, a “frame” was constructed to provide a brief introduction to ensure that the test passage could stand alone and that the test taker would have the necessary background information necessary for an accurate and fluid read of the test passage. An example of a frame is “The main character of the story has survived a plane crash. He has already spent 3 days alone in the jungles of Costa Rica. This passage picks up with the character’s attempt to find food.” Readability indices were calculated. The literary and consumer/human interest passages were computed by the Fry (1968) readability formula to have eighth grade reading difficulty; whereas, the content-based passages were computed to be at the ninth grade level of difficulty (NCDPI, 1996). The questions that the NCDPI used in its initial analyses of the passages were also used in this study.

The test items were constructed by first choosing test item writers. Twelve NC teachers were chosen by the NCDPI curriculum specialists across the state on the basis of their perceived knowledge of the curriculum and their exemplary teaching status. Classroom teachers were used to write the items because it was felt that their classroom experiences “would ensure instructional validity” (NCDPI, 1996). These item writers were given training in writing test items. The second step was to write the pool of items. Test item writers for the NC Reading Test were instructed to “address as many aspects of the curriculum as possible while developing relevant and important questions related to the passage” (NCDPI, 1996).

The third step in the initial NCDPI test construction process was to have the pool of items analyzed by curriculum specialists and classroom teachers to ensure that the test items were perceived to be “valid representations of the objectives for which they were written.” For each subject and grade level, 10–15 individuals met to review the items and passages for reading. Item reviewers learned about the revised statewide curriculum (Standard Course of Study) for the items they would be reviewing, the End of Grade (EOG) testing program, and the test development process in general. They were given a copy of “How to Review Multiple Choice Achievement Test Items” (developed by NCDPI), which includes the criteria for evaluating each item. Each of the criteria for evaluating items was discussed, along with examples of those items that met the criteria and those that did not. At least four individuals rated each of the test items being considered for the final item pool. Broad criteria raters applied to the items are as follows:

1. Conceptually, is the item a good one? Considerations for this category include the following: objective match, fair representation, lack of cultural bias, clear statement, single problem, one best answer, common context in foils, and credibility of foils.

2. Linguistically, is the item appropriate? Considerations relate to age appropriateness, correct punctuation, spelling, and grammar; lack of excess words; no stem/foil clues; and no negative in foils.

3. Is the format appropriate? Considerations include: logical order of foils; print size and type, familiar presentation style; correct mechanics and appearance; and equal length in foils.

4. Are diagrams (if used) appropriate? Considerations include necessity of the diagram, quality of the diagram, and unbiased nature of it (NCDPI, 1996, p. 17).
In addition, raters answered the following questions about each passage:

1. For which grade level is the passage appropriate?
2. For the grade level to which the passage is currently assigned, is it easy, medium, or hard?
3. Is the passage interesting to read and does it have a beginning, middle, and end?
4. Is the frame acceptable for the passage?
5. Do all the objectives fit well with the passage or should one or more not be used and substituted with another objective? Please explain.
6. Do the items adequately cover the major content of the passage? Are the most important ideas included? Please explain (NCDPI, 1996, p. 17).

The fourth step in the initial NCDPI test construction involved a team of reviewers, consisting of exemplary teachers, curriculum specialists, and test development staff (NCDPI, 1996), revising and/or rewriting items based on feedback from the curriculum specialists and classroom teachers. Frames for reading passages were also modified as necessary. The fifth step involved writing additional test items and passages as they were needed (NCDPI, 1996). These items underwent the same process as the earlier items. The sixth, and final, step consisted of the final pool of questions being reviewed by NCDPI curriculum specialists, university professors, teachers, administrators, NCDPI testing consultants, the NC Testing Commission, and others who worked together to establish the test specifications. Items were classified according to two dimensions: (a) difficulty level and (b) thinking skill level (NCDPI, 1996). Item difficulty level was determined in a series of pilot studies by giving the different items to a group of students representative of the students who would ultimately take the test. An item was classified as “not difficult” when 70% of the subjects in the pilot study answered the question correctly. An item was classified as “medium difficulty” when 50%–60% of the subjects answered it correctly. An item was considered to be “difficult” when only 20%–30% of the subjects were able to answer correctly (NCDPI, 1996).

Determination of the second dimension of test items (i.e., cognitive ability needed to answer a question) was made by having curriculum specialists, teachers, administrators, university professors, and others apply the “Dimensions of Thinking” framework (Marzano et al., 1988) to determine which specific cognitive skill would be needed to answer a question. For example, one test item might ask a student to classify several passages according to their genre, thereby, presumably, requiring the cognitive skill of organizing. Another item might ask the student to choose the best procedure for solving a problem that would, presumably, require the student to employ the cognitive skill of evaluating (NCDPI, 1996). Dimensions of Thinking was developed through the collaborative efforts of national experts in thinking skills and purports to be based on current thinking in cognitive psychology, education, and philosophy (NCDPI, 1996). This framework lists includes “core thinking skills,” which are defined as “relatively specific cognitive operations that can be considered ‘building blocks’ of thinking” (NCDPI, 1996, p. 12). The rationale given by the NCDPI for using the core thinking skills to classify items on the NC Reading Test was that “these skills have a sound basis in research and theoretical literature, are important (2) for students to have, and (3) can be taught and reinforced in school” (NCDPI, 1996, p. 12).

Field testing of items. For the first part of the field testing of the original NCDPI reading test, 90 reading passages and 716 reading items for grade levels 3–8 were field tested (NCDPI, 1996). Because one of the goals in developing the reading test was to adhere as closely as possible to the developmental English Language Arts curriculum, a means of measuring a student’s academic “growth” during the course of one school year was necessary. To establish developmental growth scales, all 10 test forms were also administered to students at the seventh and ninth grade levels. In addition, procedures for administering the test were also field tested and refined. Instructions were written; procedures for distribution were organized; and administrators were trained.

Use of the NC Reading Competency Test. In NC, at the time of the research reported here, students were required to pass the NC Competency Tests of Reading and Mathematics, also referred to as the “eighth Grade
EOG Test,” in order to receive a high school diploma. Currently, students are still expected to pass these reading and math tests, but they must pass additional tests in order to obtain a high school diploma. The NC Competency Tests were originally developed for the general population of eighth grade students; however, beginning in 1997, all English language learners, and students with disabilities, including students who are deaf, have been required to pass this test in order to receive their high school diploma—despite the fact that these populations were not considered when the state’s reading competency test was developed.

Questions about the appropriateness of the NC Reading Competency Test for deaf students. Although the NC Reading Competency Test developers used a standard procedure for determining test appropriateness, defined as being unbiased, for some subgroups of test takers (i.e., male/female, blacks/whites), no evidence was found to indicate that anyone familiar with deaf students was involved in the initial construction or to determine whether the test was appropriate for deaf students.

Purpose of the Study

The purpose of this study was to examine the appropriateness, or fairness, of the NC Reading Competency Test for students who are deaf, “using the same method” originally used to determine the appropriateness of the NC Reading Competency Test for the general population of students in NC (i.e., by using experts familiar with the curriculum and characteristics of students expected to take the test), to judge the appropriateness of test passages and items. The research question addressed in this study is Do judgments of experienced teachers of deaf students support that the NC Reading Competency Test is perceived to be fair and unbiased for deaf students? Specifically, which passages and items on Form M of the NC Reading Competency Test: (a) meet the criteria related to NCDPI item review questions and (b) address well-documented test-taking variables, related to language, content, and test format, which have been found to differentially influence reading test performance of students who are deaf and students who are hearing.1

Methods

Participants (Reviewers)

Participants, referred to as reviewers, for this study, consisted of eight experienced teachers of deaf students in eighth grade or above in two schools for deaf students in NC. This number of reviewers was twice the number (i.e., four) of reviewers used by the NCDPI in its original study of the appropriateness of passages and test items for the general population of eighth grade students in NC. Four of the reviewers in this article were hearing and four were deaf. All were white females. Six of the reviewers were from the NC School for the Deaf (NCSD) in Morganton, NC. Two of the reviewers were from the Central School for the Deaf (CNCSD) in Greensboro, NC. Attempts were made to recruit teachers from mainstream programs and from the Eastern NC School for the Deaf (ENCS). The state director of the mainstream programs was contacted but declined to have any teachers in mainstream programs involved. Teachers at the ENCSD chose not to participate in the study. For these reasons, the participants came from only two schools for deaf students. All the individuals from NCSD and CNCSD who met the qualifications were contacted and were willing to participate in the study. All reviewers were familiar with the NC Reading Competency Test (which was administered twice a year). All these reviewers had at least 3 years’ experience teaching English to deaf students, with five of them having more than 20 years’ experience teaching deaf students. The average number of years teaching deaf students was 20 years. Six out of eight teachers had master’s degrees. Five of the teachers had degrees both in English and in Deaf Education. Table 1 reflects characteristics of the reviewers.

Students in the residential schools, in which the reviewers teach, vary widely in their ability to hear and speak; however, the students tend to encounter similar struggles in learning to read, which are largely related to language and content of reading passages as well as test format variables. The average measured reading level of high school graduates in these two NC
residential schools at the time of this study paralleled that of the national average for high school graduates (i.e., third to fourth grade reading level). Whereas the majority of students had attended the different residential schools for their entire schooling, some had transferred in from public school settings when they struggled with reading and could not succeed academically. The NC reading test is given to eighth graders across the state. All students are expected to pass this test in order to get their high school diploma. Many deaf students continue to retake different forms of this test during their high school years.

Reviewers were not instructed to consider one group of deaf students (e.g., prelingually, profoundly deaf students) over another (e.g., late deafened or hard of hearing students), but comments of reviewers indicated that they were reviewing the test for prelingually, profoundly deaf students, who would most likely be impacted by language, content, or test format variables.

Instrumentation

Passages and questions. The NC Reading Competency Test consisted of 10 reading passages and 68 multiple-choice test items. Passages included: (a) four literary passages, including a poem, a short story, a mythological piece, and an autobiography, (b) four “content” passages, including one related to each of music, social studies, science, and sports, and (c) two “consumer” passages, including one “how to” passage and one “food label.” Characteristics of 10 passages, including word length, percentage of test length, and number and percentage of total test items per passage, are reflected in Table 2. The passages are arranged in this table in order of how much emphasis they receive in the school curriculum (based on NC’s Standard Course of Study) and type (i.e., literary, content, consumer). One of the primary goals of the NC curriculum, at the time of this study, for eighth grade language arts focused on students studying various literary genres, such as fiction, nonfiction, drama, and poetry (NCDPI, 1999). Because literary passages are specified in the language arts curriculum, they were listed first in this study. Students would have more experience with content passages, such as used for social studies and science classes, than with consumer passages, so the content passages were listed second and the consumer passages placed last. Within the type of passage, the passages were arranged by the number of words in passages.

Table 3 reflects difficulty level of the passages according to the Fry (1968) and Dale and Chall (1995) readability formulas. The Fry formula, used by the NCDPI in the original test construction, uses measures of sentence length and word length to calculate text difficulty. The Dale and Chall (1995) formula uses the same measures, but it also calculates the number of words in the passage that do not appear on the Dale–Chall list of 3000 familiar words. Due to

<table>
<thead>
<tr>
<th>Number</th>
<th>Years teaching deaf</th>
<th>Years teaching deaf English</th>
<th>BA degree</th>
<th>MA degree</th>
<th>CED certification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hearing teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>28</td>
<td>22</td>
<td>English and deaf education</td>
<td>Reading education</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>26</td>
<td>25</td>
<td>English</td>
<td>Deaf education</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>27</td>
<td>15</td>
<td>English and deaf education</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>English</td>
<td>Deaf education</td>
</tr>
<tr>
<td>Deaf teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>English</td>
<td>Deaf education</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>24</td>
<td>20</td>
<td>Psychology</td>
<td>Deaf education</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>12</td>
<td>8</td>
<td>Psychology/social work</td>
<td>Deaf education</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>Cinema, TV</td>
<td>None</td>
</tr>
</tbody>
</table>

*From Central NC School for the Deaf (Greensboro); other six were from NC School for the Deaf (Morganton).

bMostly elementary education.

Reading and English.
documented difficulties with vocabulary of deaf readers, it was deemed to be important for this study to have a calculation of unfamiliar words in the calculation of text difficulty. Within each category, the passages are listed in order of passage length, from the shortest passage to the longest.

*Form for reviewing passages and items.* Using questions and a format based on the initial construction of the reading test by the NCDPI, the eight reviewers in this study judged the appropriateness of the 10 passages and 68 test items of the NC Reading Competency Test. Figure 2 shows the form used by reviewers. Reviewers answered five questions for each passage and six questions for each test item.

Reviewers judged the difficulty level of each of the 10 passages as being “easy, medium, hard” for deaf students who would be required to take the test (i.e., eighth grade or above). In addition, they used a “yes–no” format to judge the following for test passages: (a) whether the passage was appropriate for students who are deaf, (b) whether the passage would be interesting to deaf students taking the test, (c) whether the passage had a distinct beginning, middle, and end, (d) whether the frame was appropriate. In addition, using a yes–no format, reviewers evaluated six

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Description of test reading passages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage</td>
<td>Passage type&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Literary</td>
<td>1685</td>
</tr>
<tr>
<td>1</td>
<td>Poem</td>
</tr>
<tr>
<td>8</td>
<td>Short story</td>
</tr>
<tr>
<td>3</td>
<td>Myth</td>
</tr>
<tr>
<td>5</td>
<td>Autobiography</td>
</tr>
<tr>
<td>Content</td>
<td>2248</td>
</tr>
<tr>
<td>7</td>
<td>Music</td>
</tr>
<tr>
<td>2</td>
<td>Social Studies</td>
</tr>
<tr>
<td>4</td>
<td>Science</td>
</tr>
<tr>
<td>10</td>
<td>Sports</td>
</tr>
<tr>
<td>Consumer</td>
<td>729</td>
</tr>
<tr>
<td>9</td>
<td>How to</td>
</tr>
<tr>
<td>6</td>
<td>Food label</td>
</tr>
<tr>
<td>Average</td>
<td>466</td>
</tr>
<tr>
<td>Total</td>
<td>4662</td>
</tr>
</tbody>
</table>

<sup>a</sup>NCDPI (1996).
<sup>b</sup>Excluding frames.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Fry and Dale–Chall readability measures of the passages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage number</td>
<td>Description</td>
</tr>
<tr>
<td>Literary</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Poem</td>
</tr>
<tr>
<td>3</td>
<td>Myth</td>
</tr>
<tr>
<td>5</td>
<td>Autobiography</td>
</tr>
<tr>
<td>8</td>
<td>Short story</td>
</tr>
<tr>
<td>Content</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Science</td>
</tr>
<tr>
<td>7</td>
<td>Music</td>
</tr>
<tr>
<td>10</td>
<td>Sports</td>
</tr>
<tr>
<td>2</td>
<td>Social Studies</td>
</tr>
<tr>
<td>Consumer</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>How to</td>
</tr>
<tr>
<td>6</td>
<td>Food label</td>
</tr>
</tbody>
</table>

<sup>a</sup>May be unreliable due to short passage length.
characteristics for each of the 68 test items. These characteristics and their relationship to the questions in the original NCDPI validation study are reflected in Table 4. If reviewers responded no to any of these questions, they were asked to provide a written explanation.

Data Analysis

Data analysis consisted of tabulating responses and listing comments from the eight teacher reviewers. Calculations included: (a) the percentage of reviewers rating a passage to be easy, medium, or hard, (b) the number and percentage of no responses to each of the four questions about passages, (c) the total number and percentage of no responses to the six questions per test items, and (d) the total sum and percentages of no responses across passages and items. Finally, (e) the total numbers for the passages and test items were combined for an overall assessment of each passage.

Figure 2  Reviewing rating form.
As reflected in Table 5, none of the reviewers rated any of the passages to be easy for deaf eighth graders. Based on reviewer judgment, six of 10 passages were indicated by more than half of the reviewers to be hard. Passage 1, a poem with dialect, was considered to be the most difficult passage, with all the reviewers rating it as hard for deaf eighth graders. Two passages (passage 4: earthquakes and passage 9: how to) had near consensus with 88% rating them hard, followed by two passages (passage 3: myth and passage 7: music) rated as hard by 75% of the reviewers, and one passage (passage 6: food label) rated as hard by a simple majority (57%) of the reviewers. More than half of the reviewers were in agreement that the remaining four passages rated a medium level of difficulty.

There are some relationships between the reviewers’ judgment of passage difficulty and particular qualities of the passages (See Table 6). Five of six passages that received an overall rating of hard were also identified by at least 50% of reviewers as being “inappropriate” for use with deaf students in the eighth grade. Each of the four passages identified by six or more teachers as being “uninteresting” to deaf students was also identified as being hard. The two passages identified by at least 50% of the reviewers as “not having a distinct beginning, middle, and ending” were also considered to be hard passages. The NCDPI indicated that they used “a majority” for ratings, which we interpreted to mean at least 50%. A 50% criterion is commonly used in curriculum alignment studies (Webb, 2002).

Table 7 reflects both the perceived overall difficulty of the passages and the negative responses to the four passage questions. This table combines both the reviewers’ perceived difficulty of passages and their responses to the NC quality indicators (i.e., the four passage questions concerning appropriateness for testing, interest to students, structure beginning, middle, end of the passage, the frame). The passages are grouped first by passage type and second by degree of difficulty within the categories (beginning with the hardest first). The passages that were judged to be most problematic, across all questions and passage categories, were passage 7 (music), passage 1 (poem), passage 6 (food labels), and passage 9 (how to). Passage 7 (music) was judged to be most problematic, with three of four questions receiving more than 50% no responses from reviewers. Two of four questions related to the remaining three passages were judged by at least 50% of the reviewers to be problematic. Six passages were judged as hard. Five were judged as not appropriate. Four were judged as not interesting. Two were judged as not having adequate structure. All the passages with a negative response on any of the quality indicators were also judged as hard. All judged as hard also had negative results on at least one quality indicator. Fully six of 10 passages received some negative judgment by at least half the reviewers.

Findings from this study indicate a fairly high level of agreement between reviewers’ judgments of relative text difficulty and those of the established, and presumably, more objective readability formulas for the literary passages and content passages. The
literary and consumer/human interest passages were computed by the readability formula used by the NCDPI to be written at an easier (i.e., eighth grade) reading level than the content-based passages, which were calculated to be written at a ninth grade level of difficulty (NCDPI, 1996). For the consumer passages, however, the reviewers’ judgments and the formulas did not correspond as well, possibly because passage 6 (food label) was judged to be highly problematic by reviewers, and there were few other consumer passages to offset the problems with passage 6. Findings for passages are presented in the following order: literary passages, content passages, and consumer passages.

**Literary passages.** For the literary passages (i.e., passages 1, 3, 5, and 8), a high degree of agreement was evident between reviewers’ judgments and readability formulas. Passage 1 was found by all indicators (reviewers, readability formulas, and % hard words) to be the most difficult of the literary passages (and of all the passages). Passages 3 and 5 were determined by the reviewers to be of intermediate difficulty (rated medium). Passage 3 was found to be the second most

<table>
<thead>
<tr>
<th>Passage number</th>
<th>Description</th>
<th>Easy n</th>
<th>Easy %</th>
<th>Medium n</th>
<th>Medium %</th>
<th>Difficult n</th>
<th>Difficult %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Poem</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>Myth</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>25</td>
<td>6</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>Autobiography</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>57</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>8</td>
<td>Short story</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>75</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Science</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>7</td>
<td>88</td>
</tr>
<tr>
<td>7</td>
<td>Music</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>25</td>
<td>6</td>
<td>75</td>
</tr>
<tr>
<td>10</td>
<td>Sports</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>57</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>Social studies</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>63</td>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>Consumer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>How to</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>7</td>
<td>88</td>
</tr>
<tr>
<td>6</td>
<td>Food label</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>43</td>
<td>4</td>
<td>57</td>
</tr>
</tbody>
</table>

*aOne person did not respond.*
difficult by the reviewers, but third most difficult, based on the formulas; whereas, passage 5 was found to be the second most difficult by the formulas but third most difficult based on the reviewers’ judgments. Passage 8 was found to be the easiest of the literary passages, based on all measures. The following provides more detailed findings about each of the four literary passages.

Passage 1, the 99-word poem, had a Fry Readability index of fifth grade and a Dale–Chall readability level of ninth to 10th grade. With so few words, the discrepancy might be related to the limitation of formulas for small sample sizes. The total number of words was 96 with a total of six sentences and an average of six words per sentence. The reviewers unanimously rated this passage as hard, supporting the Dale–Chall estimate but not the Fry estimate. This illustrates the variability in readability that can occur when different formulas are used.

Passage 3, dealing with mythology, had a readability index of sixth grade according to the Fry formula and fifth to sixth grade according to the Dale–Chall formula. The total number of words was 595 with a total of 48 sentences and an average of 12 words per sentence. Six of eight reviewers rated this passage as hard; whereas, two of the deaf evaluators rated this passage as being of medium difficulty.

Passage 5, an autobiographical account, was calculated by the Fry readability formula to have a difficulty level of sixth grade and a difficulty level of seventh to eighth grade by the Dale–Chall formula. The total number of words was 721 with a total of 53 sentences and an average of 14 words per sentence. Of seven reviewers (one did not respond), 43% (n = 3) rated this passage as hard, whereas 57% (n = 4) of the reviewers rated this passage as being of medium difficulty.

Passage 8, a short story, was indicated by the Fry formula to be have a first grade level of difficulty, whereas according to the Dale–Chall formula, it had a fourth grade of difficulty. The total number of words was 259 with a total of 41 sentences and an average of 6.1 words per sentence. Six of eight reviewers rated this passage as being of medium difficulty; two of the eight teachers rated this passage as hard.

Content passages. For the content passages (i.e., 2, 4, 7, and 10), passage 2 was rated the easiest by all indicators. Passages 4 and 7 were rated hard by all indicators. Passage 10 was considered medium difficulty by the majority of indicators. Passage 2 was perceived to be the easiest of the content passages, and the formulas rated passages 4, 7, and 10 as more difficult than passage 2. Overall, a fair amount of consistency exists, therefore, between reviewers and objective measures for the content passages. The following provides details about each of the four content passages.

Passage 4, a scientific passage about earthquakes, had a Fry Readability index of ninth grade and
a Dale–Chall index of 11th–12th grade. The total number of words was 509, with a total of 37 sentences and an average of 13.7 words per sentence. Eighty-eighth percent (seven of eight) of the reviewers rated this passage as hard; whereas, one deaf reviewer rated this passage as being of medium difficulty.

Passage 7, dealing with folk songs, had a readability index of eighth grade according to the Fry formula and an 11th–12th grade readability level according to the Dale–Chall formula. The total number of words was 342 with a total of 18 sentences and an average of 19.0 words per sentence. Six of eight reviewers rated this passage as hard; two of eight reviewers rated this passage as being of medium difficulty.

Passage 10, a passage about basketball, had a Fry Readability index of ninth grade and a Dale–Chall index of 11th–12th grade reading level. The total number of words was 735 with a total of 44 sentences and an average of 16.3 words per sentence. Of the eight reviewers, 43% rated this passage as hard, whereas 57% rated this passage as being of medium difficulty. The reviewers appear to have scored this passage as being easier than did the readability formulas.

Passage 2, dealing with the Civil War, was calculated by the Fry formula to have a sixth grade level of difficulty, whereas the Dale–Chall formula calculated the difficulty level to be fifth to sixth grade. The total number of words was 735 with a total of 44 sentences and an average of 16.3 words per sentence. Overall, this passage seemed to be fairly easy based on these reading measures. The majority of the reviewers (63%) rated this passage to be of medium difficulty; only 38% rated it as hard, and thus, the reviewers’ overall ratings seem to match the ratings from the readability formulas.

Consumer passages. For the consumer passages (i.e., 6 and 9), the reviewers’ perceptions of difficulty do not match the calculated difficulty by formulas. Both readability formulas indicated that passage 6 was more difficult than passage 9, whereas, the reviewers perceived passage 9 to be more difficult than passage 6. Based on this discrepancy, plus the fact that there were only two passages of this type, consumer passages might be difficult to judge compared to other types of passages or the pattern is less clear because of small numbers of passages.

Passage 9, a how-to passage, was calculated by the Fry formula to have a seventh grade difficulty level and a seventh to eighth grade difficulty level by the Dale–Chall index. The total number of words was 322 with a total of 14 sentences and an average of 23 words per sentence. Seven of eight reviewers rated this passage as hard.

Passage 6, about food labels, had a Fry Readability index of 11th grade and a Dale–Chall index of ninth to 10th grade. The total number of words was 309 with a total of 18 sentences and an average of 16 words per sentence. Reviewers’ perceptions of passage difficulty were split, with 57% of the reviewers rating this passage as hard and 43% rating this passage to be of medium difficulty. The readability formulas’ estimates of difficulty tended to match the reviewers’ judgments about difficulty level.

In conclusion, generally, a fairly high level of agreement was evident between reviewers’ judgments of difficulty and difficulty calculated by the readability formulas for both the literary passages and the content passages. Reviewer judgment was basically the same as the readability formulas. This finding corroborates the finding of LaSasso (1982) in a study where experienced teachers of deaf students were asked to rank the order of difficulty of four passages graded by the Fry and Dale–Chall formulas to be written between third and sixth grade reading levels. Teachers’ rankings corresponded to the rankings according to the formulas. For the consumer passages in this study, however, the reviewers’ judgments and the formulas did not match as well as they did for the literary and content passages, but it may be because of the small number of passages in that category.

Reviewers’ Judgments about the Test Items

Problematic test items were identified using a criterion of at least a 50% agreement among the eight raters, which was the criterion used in the initial NCDPI test construction. As reflected in Table 8, of the 68 test items accompanying the 10 passages on the NC Competency Test of Reading, seven test items were judged to be inappropriate (i.e., two literary items, five content items, and no consumer items). The passage with the most negative responses for the test items was
passage 7 (music) with three test items having at least 50% negative responses.

The following comments are typical of the rationale reviewers gave for indicating a no response for the test items.

**Literary passages.** For test item number 3, a hearing reviewer commented, “deaf students often have difficulty with questions presented using the negative.” Another hearing reviewer observed that there was a “negative in the foil.” A deaf reviewer indicated that “Hearing students would be more familiar” with the phrase used in the foil.

Test item number 15 had four comments, two of which mentioned critical thinking. One deaf reviewer said, “Critical thinking is involved” in answering this item. Another deaf reviewer commented, “The passage … involves critical thinking and cause/effect relationship.” A hearing reviewer said this item is inappropriate because there is “unfamiliar syntax in the question.” Another hearing reviewer, noting that performance would be affected by hearing loss, wrote, “It is unlikely that a deaf student would hear this story from a hearing parent or any adult. Storytelling by parents varies with deaf kids.”

Based on the reviewers’ comments, it is possible that the reviewers’ judgments may be based upon hearing status. That is, reviewers’ perceptions may be based on whether or not the reviewer is deaf or hearing. Deaf reviewers gave more negative marks than hearing reviewers for each of the six test item questions. For test item 3, equal numbers of deaf and hearing reviewers gave negative marks, but for test item 15, the negative marks came from three deaf reviewers and one hearing reviewer. From these comments, it appears that deaf reviewers may have some additional insight into the difficulties facing deaf readers, such as the fact that answering certain items involve the use of critical thinking skills or a recognition of cause–effect relationships, both of which involves higher level thinking skills. More research is needed to confirm this finding.

**Content passages.** Among the comments of reviewers regarding test items for passage 7 (music) are the following: A hearing reviewer chose C (performance is affected by hearing loss) because “all items refer to music.” Another hearing reviewer chose D (familiarity of content), for these items saying, “again the obvious … some hard of hearing may cope better, but that would be difficult to determine.” The third hearing reviewer picked C and F (offensive content or stereotyping), saying “the obvious.” The fourth hearing reviewer chose C and D saying, “general background knowledge— hearing would probably have some familiarity with folk songs.” Interestingly, most of the comments about music were from the four hearing reviewers.

For test item 62, one hearing reviewer observed the two different spellings of an unusual proper noun (misspelling?). She also noted that a phrase used would be problematic for deaf students. A deaf reviewer noticed the same phrase and labeled it as an “idiomatic expression.” Because of this wording, she marked B (inappropriate for testing) and D (familiarity of content). A third reviewer (second deaf) marked B and E (question format) and said, “This expression is not used/familiar to Deaf students.”

Reviewers’ comments about questions related to the content passages followed the same pattern as for the literary passages, in that deaf reviewers gave more negative marks than the hearing reviewers. The only exception to this was aspect C (performance not affected by hearing loss). Unexpectedly, hearing reviewers gave more negative responses for hearing loss.

**Consumer passages.** For the consumer passages, none of the test items received 50% or more of no
responses. Neither deaf nor hearing reviewers had problems with consumer test items, even though they did have problems with the passages.

**Differences Between Hearing and Deaf Reviewers**

When deaf and hearing reviewers’ evaluations of overall difficulty of passages are compared, it is found that they assigned almost an equal number of negative responses to questions about the passages (hearing 45% vs. deaf 55%). Differences appear, however, related to the test items. Specifically, deaf reviewers were twice as likely to assign negative responses to the test items as hearing reviewers (68%–32% or 253 no responses for the four deaf reviewers to 118 no responses for the same number of hearing reviewers).

One primary difference between deaf and hearing reviewers in their analyses of the test items is that deaf reviewers appear to recognize, more readily, the fact that certain items required the use of critical thinking and inferencing skills. The deaf reviewers made 18 comments about critical thinking skills compared to only one comment by a hearing reviewer. Another major difference between deaf and hearing reviewers related to comments concerning items that related to sound. That is, 16 comments were made by hearing reviewers versus three comments by deaf reviewers. It seems logical that the hearing reviewers would be more aware of items that were affected by sound than the deaf reviewers, although there were equal numbers of comments about sound in the passages.

**Summary of Findings**

Inspection of Tables 2 and 3 reveal that reviewer judgment ranked passage difficulty in basically the same order as the grade levels derived from the readability formulas. None of the reviewers rated any of the passages as easy for deaf eighth graders. At least half of the teachers judged six of the 10 passages to be hard. Of those six passages, however, only two had formula-calculated difficulty levels above eighth grade. Two test items in the literary passages and five items in the content passages received negative responses from at least 50% of the reviewers. When the number of negative responses regarding the test items was examined in terms of hearing and deaf reviewers, the deaf reviewers assigned more than twice as many no responses as their hearing peers.

**Implications**

Within the limitations of this study, a major finding, based on the no responses of reviewers who were all experienced teachers of deaf students, as well as their

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Summary showing perceived passage difficulty and negative judgments for the NC passage quality indicators and six item validity indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Literary</strong></td>
<td>% medium</td>
</tr>
<tr>
<td>Passage 1: poem</td>
<td>100</td>
</tr>
<tr>
<td>Passage 3: myth</td>
<td>75</td>
</tr>
<tr>
<td>Passage 5: autobiography</td>
<td>57</td>
</tr>
<tr>
<td>Passage 8: short story</td>
<td>75</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td>% medium</td>
</tr>
<tr>
<td>Passage 4: science</td>
<td>88</td>
</tr>
<tr>
<td>Passage 7: music</td>
<td>75</td>
</tr>
<tr>
<td>Passage 10: sports</td>
<td>57</td>
</tr>
<tr>
<td>Passage 2: social studies</td>
<td>63</td>
</tr>
<tr>
<td><strong>Consumer</strong></td>
<td>% medium</td>
</tr>
<tr>
<td>Passage 9: how to</td>
<td>88</td>
</tr>
<tr>
<td>Passage 6: food label</td>
<td>57</td>
</tr>
</tbody>
</table>

*Note. For difficulty level (medium, hard) and Q1–4 (NC quality indicators), the actual percentage was recorded. For the six aspects of item validity (A–F), the item number was recorded because all the responses were exactly 50%.
*Percentages are the percentage of the eight raters assigning a NO rating indicating inappropriateness.
comments about passages and items, is that many of the passages and test items on Form M of the NC Reading Competency Test are problematic and possibly inappropriate for deaf students. This finding has important implications for future NC test development, reading instruction and curriculum of deaf students, current use of test results, and research.

Implications for test development. Findings from this study suggest that it would seem prudent for NCDPI test developers who plan to use the NC Reading Competency Test with students who are deaf to consider the input of individuals who are familiar with deaf students when constructing future tests. Further, in light of the finding that deaf reviewers assigned twice as many negative responses as their hearing counterparts for the quality indicators and item validity indicators, test developers should give serious consideration to including reviewers who are deaf themselves and have the necessary qualifications pertaining to teaching reading or English language arts to students who are deaf. Having reviewers familiar with deafness involved in test construction would likely eliminate the problem observed related to sound or music as the content basis for reading passages, which are judged inappropriate for deaf students.

Subsequent to the completion of the study reported here, Martin (2005) conducted a similar study in New York, in which educators of deaf children reviewed the New York Language Arts Test. Martin was permitted to use responses of deaf students on the test for direct comparison with the reviewers’ evaluation, which was not the case for Lollis (2002). Martin found that students’ responses on the test validated the judgments of the reviewers. For example, students’ responses indicated that they struggled more on those passages perceived by teacher reviewers as being hard. This is consistent with indirect evidence related to the accuracy of teacher judgment reported by LaSasso (1982), who asked experienced teachers to rank the order of difficulty of four passages graded by the Fry and Dale–Chall formulas, which were written between third and sixth grade reading levels. LaSasso found that teachers’ rankings corresponded to the rankings according to the formulas, as well as the findings of Lollis (2002) presented above. The methods used by Lollis (2002) and Martin (2005) should be used to establish test fairness on other state-mandated tests, as well as with other special populations, such as English language learners students, to determine the appropriateness of state tests for the different populations required to take the test.

Implications for instruction and curriculum. The finding suggesting that much of the NC Reading Competency Test content was not addressed in the curriculum suggests a possible lack of alignment between the reading curriculum experienced in schools for children who are deaf and the reading test (See Webb, 2002 for a discussion of issues related to curriculum and test alignment). Readers may question the quality of the teachers or the schools in this study; however, a careful examination of the curriculum of NC for students in eighth grade and above will show that only a minor portion of the language arts curriculum deals with specific reading skills for students still learning to read (NCDPI, 2004). Nonetheless, it is obvious that the general assumption is that students have finished learning to read by the end of eighth grade, which is not the case for many deaf students.

Questions concerning the quantity and quality of explicit reading instruction in the schools serving deaf students may also be raised from findings in this study. At the time of this study, the NC schools for deaf children were expected to follow the statewide curriculum and, therefore, did not have in place a curriculum dealing specifically with deaf students and their special language and reading needs. Also, educators of children and youth who are deaf are often not given the same access to teacher training and current research in the area of reading that their counterparts in regular education would have. As LaSasso (1999) notes, reading curricula for students who are deaf need to explicitly address test-taking skills because many deaf students lack specific test-taking skills. Deaf students, at all instructional levels, need regular exposure to and systematic practice in taking the different kinds of tests that their hearing peers take.

The fact that many comments were made by reviewers knowledgeable in teaching students who are deaf, regarding critical thinking skills, was
interesting in light of the fact that inferencing and critical thinking skills were not addressed as part of the format of the test. Moreover, reviewers’ judgments that many of the questions on the NC Competency Test require critical thinking, including inferencing, which they implied would be skills lacking among some deaf students, supports the curricular need for explicit instruction in higher level thinking skills at all instructional levels, such as Feuerstein’s Instrumental Enrichment Program (Feuerstein, Rand, Hoffman, & Miller, 1980). Inferencing and other critical thinking skills are vital for fluent reading and test-taking abilities. Although some instruction for deaf students has successfully incorporated thinking skills curriculum into the schools, a greater priority needs to be given to this aspect of instruction. Given these needs, there is ample room for improvement when it comes to deaf students’ opportunity to learn. Both curriculum and instruction, at the time of the study, may have been inadequate relative to the expectations embedded in the NC Reading Competency Test.

**Implications for use of test results for deaf students.** Because of the high-stakes nature of the NC Reading Competency Test, it is incumbent upon teachers and program administrators to understand the limitations of the test for deaf students and advocate for alternative types of reading assessment, including portfolios. The “Standards for Educational and Psychological Testing” emphasize that students’ futures should not be decided on the basis of a single test, even with numerous administrations of that one test (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). The use of multiple measures, such as portfolios and other alternative assessments, would permit a more accurate assessment of deaf students’ reading abilities, thereby lessening the possibility of test bias occurring due to lack of exposure to content or language. Cawthon (2007) conducted a national survey of accommodations and alternate assessments for students who are deaf or hard of hearing in the United States and found that the most often reported types of alternative assessments used with students who are deaf are out-of-level testing, work samples, and portfolios (p. 337).

The finding that reviewers in this study found many passages and items on Form M to be problematic also suggests that care should be exercised by all when interpreting a NC deaf student’s performance on the NC Reading Competency Test. It is not clear, however, whether the problem is failure to measure the intended construct; that is, it is possible that the results are an artifact of poor test design or that students are having an inadequate opportunity to learn prior to being required to take the test.

**Universal design for assessment (UDA)** (Thompson, Johnstone, & Thurlow, 2002) offers a possible remedy for some of the problems found in our study of the appropriateness of the NC reading assessment for students who are deaf. UDA is a general framework for improving the design, and therefore accessibility, comprehensibility, and validity of tests (Johnstone, 2003). Thompson and Thurlow (2002) and Thompson et al. (2002) describe seven elements of UDA, which can be used to examine tests and their level of accessibility. Two of these elements relating to this article include: (a) inclusive assessment population and (b) accessible, nonbiased items. According to the UDA, different types of students who will be taking the test must be included in the design and field testing procedures. In the context of the study described here, students who are deaf should have been included in the field testing of the NC reading test. Further, test items need to be reviewed in advance for advantages or disadvantages in presentation or content for the different groups of test-takers, which could invalidate the item’s contribution to a test score. In the context of NC reading test, passages and test items should have been examined by individuals who are familiar with the reading abilities of students who are deaf who would be taking the test. Thompson et al. (2002, p. 6) note that “applying principles of universal design will improve the confidence we can have in test performance for all students who participate in the assessment.”

**Implications for research.** Findings from this article also have implications for future research. At least two types of analyses of student responses to test items on standardized reading tests are needed. First, the finding that reviewers in this study found so many test
items to be problematic for deaf students on the NC Reading Competency Test suggests that some of test items on the test may be biased against students who are deaf. Test bias could be investigated via the same statistical procedure (i.e., the Mantel–Haenszel DIF procedure), which was used by the NCDPI in the original test construction in order to determine whether individual test items were for two groups (male–female and black–white). The DIF procedure, which is typically used in initial test construction to reject items that are biased for one group over another, statistically compares performance of two groups on each question while controlling for comparable overall test performance. Findings from the study reported here would be strengthened if a relationship could be found between (a) test items found to have large Mantel–Haenszel DIF indices for students who are deaf and (b) test items that were identified by reviewers in this study as being problematic for deaf students. As noted earlier, we had originally planned to perform a DIF analysis on test items; however, the administrator responsible for the schools for students who are deaf was unwilling to let us administer the test to collect more test data for deaf students.

Second, it would also be helpful to analyze distractor responses to items to determine whether students who are deaf exhibit a different distribution of responses on specific items than hearing students. Differences in distractor response selection might provide additional evidence that students who are deaf are using different test-taking strategies, such as visual matching (LaSasso, 1985) or position of distractors (Trybus & Buchanan, 1973) to select responses to questions.

Research is also needed to determine the optimum qualifications of test reviewers. In this article (Lollis, 2002), discrepancies were found in ratings assigned by deaf and hearing reviewers. In the absence of corroborating statistical data, such as that could be provided by the Mantel–Haenszel procedure, it cannot be determined which group is more accurate. It could be that reviewers who are deaf are more apt to be strict about the appropriateness of test passages and items for deaf students. It could also be that these differences indicate an understanding on the part of deaf reviewers that hearing reviewers may not have. On the other hand, hearing reviewers are more likely to identify which items are affected by sound, and given the review procedures employed, these items are likely to be changed on high-stakes tests due to an identified bias. Further research is needed to determine the meaning and value of the different tendencies of deaf and hearing reviewers observed in this study.

Conclusions

The issues surrounding the assessment of reading for students in the general population are at best complex, and they become even more complicated when considering students with special circumstances, such as students who are English language learners and students who are deaf. The study reported here, the first to examine the appropriateness of a state-mandated reading test for students who are deaf, addressed one aspect of test appropriateness: test bias or test fairness. Our study employed the same procedure used by NC to establish the fairness of the original NCDPI reading test: Experts judged passages and items for students who would be expected to take the test. Our reviewers, unlike the original reviewers, were familiar with students who are deaf who would be expected to take that test. Findings from this study suggest that had our reviewers been involved in the original construction of the test, this test would have looked much different than it did and there could be greater confidence in making high-stakes decisions about deaf students based on their performance on this test. Numerous problems related to both passages and test items for deaf students were noted by the experts in our study who are familiar with deafness. These findings have important implications for test developers as well as those who make high-stakes decisions about students, including students who are deaf, based on their test performance. As Scheuneman and Oakland (1998, p. 85) note, “standardized test scores will likely remain an important component of high-stakes decision making in education for the foreseeable future. The challenge will be to interpret scores appropriately for all examinees. Meeting this challenge will require mindful awareness of the factors that may cause scores to be lower than they should be for some individuals.” Constructors of standardized reading tests to be used with
students who are deaf and those responsible for making high-stakes decisions about students who are deaf from test results must be mindful of the test-taking research related to students who are deaf.

Note

1. In this study, “difficulty” is assessed for passages only. Originally, we planned to assess test item difficulty using the DIF protocol (Mantel & Haenszel, 1959). We obtained test data for hearing students and permission to use it in this study; however, the administrator responsible for the schools for students who are deaf was unwilling to let us use actual student data for deaf students.

References


Received December 10, 2007; revisions received March 31, 2008; accepted April 14, 2008.