

Introduction

Within the context of standards-based educational systems, states are using large scale reading assessments to help ensure that all children have the opportunity to learn essential knowledge and skills. Students participating in assessments today are more diverse than they were just a decade ago, and participation rates of students with disabilities, in particular, have increased dramatically. These increases mean that states and test developers are obligated to ensure that their assessments are accessible so that all students can show what they know and can do. An accessible assessment is one that reveals the targeted knowledge and skills of all students, including students whose characteristics create barriers to accurate measurement using traditional reading assessments.

It is particularly challenging for some students with disabilities to show their knowledge and skills on reading assessments, especially when their disabilities affect reading. Students with learning or intellectual disabilities or speech-language, hearing, or visual impairments are often identified as having learning characteristics that impact the reading process. For example, difficulties with print reading and comprehension of text are problems commonly associated with learning disabilities. Students with intellectual disabilities often have poor short term memory, low-level meta-cognitive skills, difficulty with expressive language, and poor use of logic and organization.

The challenge for developers of accessible reading assessments is to develop assessments that measure only those student characteristics that are essential parts of the reading proficiency the test intends to measure, and not those characteristics that could be related to the student's disability. This attempt to untangle the effects that various student characteristics have on how a student performs requires thoughtful attention to the validity of the inferences made from assessments. The importance of clearly defining what the test is intended to measure is evident in most discussions of test validity. Unless the construct the test is intended to measure (the skills, abilities, knowledge, etc.) and the interpretations the test is intended to support are clearly specified, it is impossible to determine the extent to which the test is measuring the intended construct and the extent to which it is measuring some aspect of the student's disability.

It is possible that assessments that lead to better interpretations about the reading proficiencies of some students with disabilities are ones that have been changed in ways that are relevant to the construct of reading as it is typically understood. A student may have some characteristic that interferes with one aspect of the reading process and yet be quite proficient on other aspects. For example, students may be proficient in comprehending the meaning of a paragraph in a story, but may not be able to decode the words in the paragraph. In this case, the student might need access to the written word through a screen reader or a read aloud accommodation. If inability to do one aspect of a task prevents a student from showing the reading proficiencies he or she does have, the assessment fails to provide information that supports accurate interpretations about what the student can do.

Accessible assessments move beyond merely providing a way for students to participate in assessments. They provide a means for determining whether the knowledge and skills of each student meet standards-based criteria. This is not to say that accessible assessments are designed to measure whatever knowledge and skills a student happens to have. Rather, they measure the same knowledge and skills at the same level as traditional large scale reading assessments. Accessibility does not entail measuring different knowledge and skills for students with disabilities from what would be measured for peers without disabilities.

The National Accessible Reading Assessment Projects (NARAP) have been conducting research to identify ways to increase the accessibility of reading assessments. This document is the culmination of one of NARAP's goals: to develop evidence-based principles for making large scale assessments of reading proficiency more accessible for students who have disabilities that affect reading, while maintaining a high level of validity for all students taking the assessments. Some of the principles clarify and underscore the importance of well-accepted and widely used practices in designing reading assessments. Other principles have been developed from theory to respond to the needs of specific groups of students.

The principles are to be viewed as a whole, representing a coherent and integrated approach to accessibility. They provide a vision of accessible reading assessments. It is not appropriate to meet one or two principles when designing an assessment and assume that the assessment is accessible. Meeting all principles, however, also does not guarantee accessibility of the assessment. Research is needed to document both the validity of the test scores and accessibility of the assessment.

Audience. This document was written primarily for personnel in state assessment offices and for test developers of regular large scale reading assessments used for accountability purposes. Other audiences also may find the document to be of interest and useful for other types of assessments.

State assessment personnel and test developers can use the principles to develop reading assessments that are as accessible as possible. For those states that already have assessments in place, these principles can be used to determine how accessible existing assessments are, so that when these assessments are revised in the future, improving accessibility can be included among the revisions. The principles can also be used to help evaluate new assessments, such as online tests, that may be under consideration.

Appropriate Use. The principles address regular large scale assessments of reading. Specifically, they are focused on reading assessments of content standards based on grade-level achievement standards used for accountability purposes (either school or student accountability).

The design of the reading portion of normative assessments included in state assessment programs and the reading component of English language proficiency assessments can

be guided by the principles included in this document. The principles focus on grade-level content and achievement and illuminate how standards-based assessments of reading might be adjusted to reveal the reading capabilities of all students, particularly those with disabilities. This focus clarifies and emphasizes the point that increasing accessibility does not imply changing the content or providing lower level or easier materials.

Because of the emphasis on grade-level content and achievement standards, it follows that the accessibility principles apply to alternate assessments based on grade-level achievement standards. It also follows that there are no limitations on the students for whom they may be relevant. With improved accessibility, regular assessments might be appropriate for students who, without increased accessibility, would have participated in an alternate assessment based either on modified or alternate achievement standards.

Although the principles presented here center on reading assessments for students with disabilities, they have broader applicability. Some of the principles may also be used for other content areas, such as mathematics and science. They are also relevant for all students, not only those with disabilities. When developers of an assessment are careful to define the target of measurement so that extraneous factors will not be inadvertently measured, there is benefit to all students. In some cases, such as when specific access tools are incorporated into the assessment itself, there is an extra burden on the developer to ensure that the validity of results – the extent to which the results provide an accurate indication of the student’s knowledge and skills on what the assessment was designed to measure – are not compromised. Some may see designing accessible assessments as a balancing act to incorporate both accessibility and validity, when in reality accessibility creates results that more accurately reflect a student’s knowledge and skills in relation to the target content, thus producing greater validity of the results.

Development of the Principles. The development of this document was accomplished through a process of first identifying principles drawn from research, existing standards, and the consensus of experts in educational assessment and then subjecting the principles to a series of reviews. Annotated references for the support of the principles are provided in the appendix. At several points, NARAP staff sought reviews of the principles’ format and content. This occurred informally during the development process as staff in the three projects within NARAP refined the principles and sent them to members of their technical advisory committees for reactions. The review process also occurred formally during three events.

The first event was the NARAP Principles Committee meeting in February 2008, during which committee members provided formal input on the principles. This committee was a stakeholder group of reading experts, state assessment and special education personnel, teachers, parents, and students.

In the second event, the principles were presented during a workshop at the Association for Test Publishers conference in March 2008. At this workshop, attendees applied

revised principles to existing assessments or to the plans for new assessments.

The principles were also presented during the third event, the Council of Chief State School Officers National Conference on Student Assessment. The conference was attended by large numbers of state assessment personnel and test developers, the target audiences for this document. A formal interactive session was used to garner input from these stakeholders.

Overview of the Principles. The focus of this document is on accessibility *principles*. The term *standard* is not used here to define the rules that emerged from the work of the projects. *Standard* is a term used by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) in the *Standards for Educational and Psychological Testing*. The principles in this document are consistent with but not as broad as the AERA, APA, and NCME *Standards*. We chose the word *principle* because of the aspirational and visionary notion that it implies. The term *standard* suggests something that must be met or adhered to. Our intent was to be more goal-oriented, rather than directive. Here principles are presented as rules that define the overarching goals to achieve accessibility.

Five principles provide the frame for accessibility for reading assessments. Each of these principles is supported by specific guidelines that address the implementation of the principles. The five principles are:

- Reading assessments are accessible to all students in the testing population, including students with disabilities.
- Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence, and attuned to accessibility concerns.
- Reading assessments are developed with accessibility as a goal throughout rigorous and well-documented test design, development, and implementation procedures.
- Reading assessments reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student's proficiencies.
- Reporting of reading assessment results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results.

This document describes the principles and provides a rationale for each. The guidelines are explained and examples are provided.

Principle 1: Reading assessments are accessible to all students in the testing population, including students with disabilities.

Rationale. Large scale reading assessments must show whether students have developed the knowledge and skills specified in reading content standards well enough to perform at the level specified by achievement standards. Students with disabilities present particular challenges because their disabilities sometimes interfere with their performance on the assessments. The accessibility of these assessments needs to be carefully considered and planned. This involves understanding the many sensory, physical, and cognitive disabilities students can have that create specific barriers to demonstrating reading knowledge and skills. Increasing accessibility also involves applying principles of universal design, using new technologies and other approaches as appropriate, and employing assessment accommodations when needed to obtain valid measures.

Guideline 1-A. Understand and account for the range of student characteristics and experiences that impact reading when designing reading assessments.

Challenges with short-term memory, organization, attention, and other skills not intended to be assessed by reading assessments should be considered when designing or revising an assessment. The validity of the results from the assessment is threatened when these characteristics are not considered during initial design or during revision. By understanding and accounting for the range of student characteristics and experiences, including disabilities and other characteristics such as cultural and language background, it is possible to enhance an assessment's accessibility from the beginning. This does not entail changing what the assessment is designed to measure. It simply means ensuring that the characteristics of the students are not inadvertently measured because they are not known or understood.

Example. Test developers should ensure that their item writers know about the range of characteristics of the students who will be taking the items they develop. Understanding the implications of such characteristics as short-term memory limitations and their relation to the intent of the measurement, which most often is not to test the student's memory capabilities, aids item developers in producing items that have low memory load requirements. An example of a high memory load item is one that requires the student to retain specific information about how to answer items (e.g., pick all answers that are correct) at the beginning of the test (before a passage and items are presented), rather than with the questions. An item that has lower memory load is one that provides specific information about how to answer items at the point the student is about to answer questions.

Guideline 1-B. Begin the development of reading assessments by applying the elements of universal design.

Universally designed assessments have been defined as grade-level assessments built

from the ground up and continually refined to be accessible and valid for the greatest number of students. Specific elements and considerations for implementing universal design principles have been delineated, with research evidence accumulating about their effects. Although most universal design research has been conducted on mathematics assessments, research in the area of reading is being conducted and evidence gathered.

Example. Among the elements that underlie universally designed assessments are precisely defined constructs; nonbiased items; and simple, clear, and intuitive instructions and procedures. These should be applied to reading assessments. Researchers now have provided clear support, for example, for the process of item and test review using methodologies such as differential item functioning data analysis and think-aloud analyses or cognitive labs to ensure that items are not biased. An example of an item that is universally designed is one that was originally conceived by writers who had been trained on the characteristics of both students with disabilities and items that are non-biased, simple, and clear. The item will have been checked through item analysis procedures and student cognitive labs.

Guideline 1-C. Use technologies and other evidence-based approaches to provide all students with a variety of assessment options within a similar testing experience.

Assessments that seek to be as accessible as possible to the widest range of students use a variety of approaches to ensure that assessments produce similar experiences for all students in ways that produce valid results. To obtain valid results for students with diverse characteristics, the similar experiences must be more flexible and broader than they generally are now. Approaches may include allowing students to select their own passages from those of equivalent difficulty or using technologies so that all students have multiple means of information retrieval and multiple means to express themselves (e.g., voice-activated technology; screen readers).

Example. Computer-based testing that includes options for students to use tools that do not compromise the construct tested is an example of this approach. Whether a student has a disability, is learning English, or simply needs assistance during the test on factors that are not focused on the construct, the result is a seamless provision of access tools to all students. For example, students may have the option of having words pronounced for them by the computer, making their responses via keyboard or voice-activation software, or using a host of other approaches that are part of the programming for the assessment. Computer-based testing is also an effective means of providing assistive technology that meets the individualized needs of students with disabilities. For example, the appropriate use of assistive technology such as screen readers, reading pens, or other text-to-speech conversions should be considered when assessing some aspects of a student's reading proficiency.

Using technology also enables the assessment to measure reading comprehension and foundational skills separately. The scores of students who have difficulty developing skills such as word recognition and phonological decoding, for example, are negatively affected

when the assessment requires that these skills be measured as a part of comprehension skills. Students may have a better opportunity to demonstrate their comprehension skills through the use of assistive technology.

Guideline 1-D. Document decisions that are made to make tests more accessible, and monitor the effects for different groups of students.

It is possible that some of the adjustments that make assessments more accessible in general will create challenges for specific students who have certain characteristics that need to be considered. For example, adding certain graphical elements may create barriers for students who are blind or have low vision. Documentation needs to be kept on each decision made about accessibility, so that potential implications for specific groups of students can be reviewed and studied in the future.

Example. Teachers or test administrators can note the difficulties that students have during the assessment when the accessibility features are present. They can note also whether the difficulties arise only for certain student groups. Documentation will reveal when common difficulties are noted across more than one test administrator or teacher. Problems can be solved collaboratively around the pattern of difficulties and issues addressed.

Principle 2: Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence, and attuned to accessibility concerns.

Rationale. To be effective, large scale reading assessments should be based on a clear and coherent definition of reading. Definitions of reading describe the complex *processes* that students engage in when they read. An example is the NAEP 2009 Reading Framework definition, which defines reading as follows: “Reading is an active and complex process that involves understanding written text; developing and interpreting meaning; and using meaning as appropriate to type of text, purpose, and situation.” NAEP draws upon what we know from cognitive, perceptual, and linguistic theoretical models that explain reading processes and the underlying constructs of reading. Current definitions also acknowledge that reading is a social and cultural activity. This means that students’ understandings are shaped by their interactions with others and who these young people are as cultural beings. In addition, definitions are shaped by broader contextual dimensions or affective frameworks such as students’ interests, attitudes, and motivations.

Definitions of reading and their underlying constructs, used as the basis for reading assessments, should be complemented by coherent and conceptually sound models of *literacy learning*, including attention to the following: (a) *the reader*, specifically where readers are located in the developmental continuum, what they should be able to do to demonstrate reading proficiency, and how they enact their feelings of self-efficacy toward the act of reading; (b) *the text*, including the range of genres or text types that students encounter, the complexities and challenges posed by each passage, and the support offered within a text including such features as headings, pictures and graphs, and other features or organizational structures; and (c) *the context*, including the various purposes that are set or selected by a reader (e.g., reading for information or reading for enjoyment), and the tasks students are asked to engage in within the overall testing situation, such as the questions posed after the passage and other ways students are invited to show what they can do on the assessment. These three features within the model of literacy learning determine the extent to which an assessment is designed that either supports or inadvertently places students at risk of not being able to demonstrate what they are capable of as readers.

Guideline 2-A. State standards are grounded in an evidence-based definition of reading.

To accurately indicate students’ reading abilities at different grade levels, state standards for each grade level should be grounded on an evidence-based definition of reading. This means that standards would require individuals to: (a) interact with appropriate grade-level written texts (e.g., literary and informational); (b) adapt reading behaviors (rate of reading; skills and strategies applied) based on a variety of purposes and situations; and (c) interpret the meanings of these texts using a range of cognitive targets. Cognitive targets are behaviors and skills appropriate for a particular grade level (e.g., decoding

and other word recognition skills; comprehension skills such as locate/recall, integrate/interpret, and critique/evaluate).

Because learning to read is a developmental process for students, appropriate cognitive targets would be developed for each grade level and subsequent grade levels would build on these cognitive targets in an incremental manner based on research on how reading ability develops (e.g., task sophistication and text difficulty). State standards should be coherent and consistent in the way reading is conceived and articulated through standards statements across the grade levels.

Example. A cognitive target may be linked to an evidenced-based definition of reading in the following way: When students are asked to infer (cognitive target) from a text they have selected to read to learn about wolves, they are required to draw upon their schemata and sort out appropriate ideas using prior knowledge and features of the text (e.g., drawing upon cognitive models and perceptual and linguistic frameworks). They are also reading a text they have chosen, for a specific purpose (e.g., grounded on an affective theoretical framework).

Guideline 2-B. Design reading tests to allow for individual measurement of the components reflected in state standards.

All students deserve the opportunity to show their developing proficiency as readers. Some students may be able to demonstrate skills on a range of important reading standards even though they may not be proficient on all components of reading proficiency.¹ For example, students with disabilities may have weaknesses in decoding or reading fluency that interfere with comprehension and vocabulary knowledge. In contrast, students with some learning disabilities or autism spectrum disorder may have skills in decoding and fluency but not comprehension. For students who demonstrate these and other diverse proficiency profiles, it might be possible to obtain a better estimate of reading proficiency if skills such as decoding or constructing meaning are measured separately as part of a comprehensive test of reading proficiency.

Example. A number of options can be employed to ensure that a reading assessment provides a complete picture of a student's reading proficiency as defined by the state's standards and test blueprints. One option is to design a test that allows for the separation of decoding and the skills of constructing meaning so that poor performance on one does not negate performance on the other. This could include separate direct measures of decoding or fluency or designating test items as decoding with comprehension and others as comprehension only. A second option might be to provide a branched test that provides an overall measure of reading proficiency and, depending on performance on that measure, branches students into modules or tasks that focus on grade-level components reflecting the state's grade-level standards. Word recognition and fluency

¹ Reading proficiency is influenced by the following components of reading: phonemic awareness, decoding, fluency, background information and vocabulary, strategies to construct meaning from print, and motivation to read (see *Defining Reading Proficiency for Accessible Large Scale Assessments: Some Guiding Principles and Issues*, <http://www.narap.info/publications/reports/definingreadingprof.htm>).

typically enable comprehension, and knowing an individual's performance in these subcomponents might help understand the student's overall proficiency.

Guideline 2-C. Use criteria to select texts that represent different genres and promote the use of interesting passages that are likely to engage all students.

Reading frameworks and definitions that account for affective aspects of reading provide support for the idea that test developers should select reading passages because they are interesting and engaging to students. Research on students' interests provides an evidence base for selecting topics. In addition, students bring a broad range of background knowledge to an assessment. Criteria for selecting interesting passages should include attention to a range of topics, including content that appeals to both genders and multiple ethnicities and that considers the varying interests of students at different grade levels.

Grade-appropriate texts should be selected using specified criteria, including the use of: (a) passages that provide depth to the concepts and ideas presented, (b) texts that are of high quality, (c) texts free from bias that support fair assessment of a diverse population, and (d) passages selected from a full range of genres that students typically encounter across the grade levels. Literary texts that include fiction, literary nonfiction, and poetry should be included on assessments as well as informational texts (e.g., exposition, argumentation, and persuasive writing) and procedural texts and documents. Various types of texts should be included on reading assessments in proportions appropriate for the grade level of the students. Texts selected for inclusion on reading assessments should also be considerate texts, meaning they should be well-written, with coherent structures that provide a strong organizational framework for ideas and that support the information presented.

Example. One criterion states may use to create accessible reading assessments is to select passages that represent the kinds of texts and topics that students read during quality reading instruction and texts they read both in and out of school. This criterion promotes the idea that students may engage more deeply with interesting texts that are well-written and they may persevere through the passages and complete all items. Another criterion is to promote the use of narrative, expository, and poetic texts on assessments (although not all forms of these broad genres could be used in a single assessment). Providing students with opportunities to show that they can read and comprehend texts that are structured in a variety of ways (e.g., comparing and contrasting information vs. a story structure with characters and plot) is important to providing a complete and valid measure of reading ability.

Guideline 2-D. Use criteria to determine the visual elements that should be included within texts while avoiding distracting elements that impact students' comprehension.

The use of intact texts or excerpts of intact texts with their original features on large-scale

reading assessments promotes ecological validity. Whenever possible (and if appropriate for the target population and purpose of the test), passages used in assessments should retain the characteristics of the original texts, including the visual elements (e.g., maps, charts, pictures) and the layout characteristics, such as the juxtaposition of text and visual elements in the original text.

Criteria should be developed that outline when visual elements are retained or removed and why. These criteria are critical because research studies differ in their findings on this matter. Some studies indicate that visual elements enhance students' reading comprehension and engagement with texts. In contrast, some studies indicate that the use of irrelevant pictures, graphics, or formatting features may be particularly distracting, and even possibly misleading, for students with learning disabilities. By irrelevant we mean pictures that are tangential or even contradictory to key ideas in the passage, or artwork that is included to add interest to the page but does not link with the print message. In addition, visual elements should not be used if they provide key concepts that are not described in the printed text, but that are assessed; this practice might disadvantage students with visual impairments.

Example. A criterion states may use to determine whether a visual element should be retained is to determine if the visual is important to understanding the concepts discussed in the passage. For example, if the author has included a picture or map to help readers understand a concept that is key to the passage or explained in the text, then the visual element should be retained. If such a visual is retained, it is very important that care be taken to provide an accurate description of the visual element that can be accessed by an examinee with a visual impairment. A second criterion states may use to retain a visual element is if it appears to enhance a reader's motivation. Examples of this include the fact that many authors of passages used in assessments secure photos or artistic drawings to help readers connect better with characters or settings in narrative texts or events or concepts in expository texts. Authors also use visual elements to help break up the text for the reader, thus reducing some anxiety about reading long sections of text, helping readers sustain their efforts. In situations where visual elements are used for the purpose of motivation or reduction of anxiety, it is important that the needs of examinees who are blind or visually impaired be considered. If the introduction of a visual element, or retention of a visual with the original text will make the task more difficult for these students than for students without a disability, an alternative to the visual should be considered.

Guideline 2-E. Present reading tasks that students perceive as worthwhile and that enable them to be self-efficacious as learners.

Reading frameworks and definitions that account for affective aspects of reading support the idea that the tasks associated with passages used in large-scale assessments should be selected or designed to ensure that students perceive them as realistic, worthwhile, and doable. Even if students are competent at an activity they may not engage in it if they do not find a purpose for doing so. Students who find reading tasks to be purposeful,

interesting, and challenging (but not too difficult) engage with these tasks and tend to persevere, even when they encounter difficult sections in a text. Research indicates that students—including those with disabilities—who perceive reading tasks that they encounter in school as worthwhile, expect to be successful readers. A positive stance toward reading assessments is contingent on providing students with interesting tasks.

Tasks or items on reading assessments must be carefully designed to measure grade-level cognitive targets that are appropriate to the problems or ideas inherent in a text. The items also need to be designed with the students' motivation and engagement in mind. The tasks should be interesting and challenging, yet achievable. Tasks should require students to use flexibly the skills and strategies they have developed, as required by the demands of the text and task. This practice promotes self-efficacious readers who believe that they have the ability to perform on tasks they have yet to face. In addition, tasks should match those that students encounter when engaged in quality reading instruction as well as out-of-school reading.

Example. Test developers could employ a criterion that promotes an expanded notion of tasks on reading assessments so that students find activities purposeful. For example, students could be asked to read to solve a mystery or summarize to write a story ending or read to determine the next steps needed to conduct an experiment. A second criterion might promote that idea of offering students some choices among several texts of similar genres and difficulty to read and respond to. Offering choice to students represents challenges and there are controversies surrounding this idea in the research literature. The benefits may be that some choice in the testing situation can provide students with a sense of control and autonomy and may enhance a student's performance on a reading assessment. A third criterion is that, whenever possible, items should be developed that move beyond one way of assessing students' reading abilities (e.g., multiple-choice items) to a range of accessible tasks that allow all students to show their proficiencies.

Guideline 2-F. Ensure that test blueprints are aligned with the state standards.

When large-scale reading tests are designed, blueprints and test specifications also need to be created for test developers to follow that outline the intent of the test, the requirements for the selection of passages and items, the design and layout of the test, and the coherent features of the test. Developers should have a firm understanding of the concepts and definition of reading guiding the development of an accessible assessment because this will impact the texts selected for the assessment and the reading abilities assessed at each grade level.

Example. Test developers need to understand the intent of grade-level cognitive targets and create items that match the text and the intended reading task or goal. Examples of these items should be provided. To accomplish this goal, state department of education personnel should require that a diverse committee be assembled that includes test developers, teachers, reading experts, and those with expertise in all different disability areas.

Principle 3: Reading assessments are developed with accessibility as a goal throughout rigorous and well-documented test design, development, and implementation procedures.

Rationale. Employing test design and development procedures that are recognized as good practice for creating any assessment is especially important for ensuring that reading tests are accessible. Accessibility concerns highlight the importance of attending to the needs of all test takers throughout standard test design and development procedures. The ideal is to plan for accessibility from the outset, but there are also steps that can be taken when new items or versions are developed for existing reading tests, and existing reading assessments can be retrofitted to be more accessible by using accommodations.

The assessment literature describes broadly accepted procedures for developing evidence regarding how well certain tasks and methods for observing and reporting on test takers' responses to the tasks can be used as a basis for making accurate and fair inferences. The process starts with establishing a clear understanding of what the assessment is intended to accomplish for whom. An important part of that understanding is consideration of possible unintended as well as intended consequences that may result from use of the assessment. Tasks and observation and reporting methods are selected or developed that seem likely to inform the inferences that need to be made. These tasks and observation and reporting methods are tried out to determine how well they inform the inferences. Adjustments may be made and further information gathered about the assessment's effectiveness. Information is collected during this entire process and reported to potential test users so they can make an informed decision as to how well the assessment meets their needs.

Guideline 3-A. Initial test design considers the characteristics of all test takers.

All intended populations of test takers should be stated in the initial specifications for the test. Stakeholders who can provide insight into test takers' characteristics should be consulted. All subsequent decisions about test content, format, and the like should be made with all student groups in mind. This includes planning for how the assessment results might be used. Consideration of all student groups extends even to how the constructs that the reading achievement assessment will measure are defined.

Example. An example of considering the characteristics of all test takers when defining the construct is asking whether a reading assessment is intended to measure both comprehension and foundational skills. For reading achievement assessments, thought should be given to defining the construct so that it is possible to measure the different components of reading separately (for example, decoding, word recognition, comprehension). If the components of reading achievement are measured separately, it may be possible to provide accommodations for one component of the reading construct and to assess another component without an accommodation. This could be done at the item or test section level. An example of considering the needs of all test takers in

light of different test purposes might contrast the implications of using certain testing accommodations on tests that are part of a college admissions requirement with using those accommodations on tests that are part of a state's school accountability system.

Guideline 3-B. Item development and evaluation considers the characteristics of all test takers.

Procedures used to develop and evaluate assessment items and tasks should be clearly specified and the procedures should be carefully followed and documented. Item writer and task developers should be well-trained regarding the varying needs of test takers and should have ready access to specialists who can clarify population needs. The content and format of the items or tasks may be modified, to some extent, to increase accessibility for all subgroups of the population. However, these items or tasks must be clearly aligned with grade-level standards and should not be modified to provide an academically less demanding task. It may be the case that some subgroups of the intended population for the reading assessment are not receiving instruction that is adequately preparing them for a grade-level assessment; if this is the case, issues regarding the instruction of these subgroups should be addressed, but the assessment should remain focused on grade-level achievement.

Example. One example of the application of this guideline might occur during the item development process when draft test items are evaluated by a diverse group of reviewers who include content experts, members of groups that comprise the intended population for the assessment, and experts in particular areas that may influence how a test taker's disability might interact with the design and administration of the assessment items or tasks. Another example is that once items have passed this initial review and revision, they are typically field-tested with samples large enough to provide statistical data that is adequate for evaluating the quality of the item or task. Item or task tryout samples should include all groups of the intended test population. Every effort should be made to carry out analyses, such as differential item functioning, for each student group, including each disability category. Items flagged by such statistical analyses should be scrutinized further to see whether the item has characteristics that require its removal from the item pool. When the student group is too small to permit meaningful statistical analyses, other types of research such as think-aloud studies should be conducted to evaluate the appropriateness of the items or tasks.

Guideline 3-C. Test assembly and evaluation considers the characteristics of all test takers.

An assessment is more than the sum of its parts. Using only items that have individually been found to be accessible does not guarantee that the resulting test will be accessible. Steps need to be taken when assembling items or tasks into a test to ensure that the entire test remains accessible. Accessibility obstacles can arise in numerous places and ways when tests are being assembled. For tests that will be used repeatedly, ongoing evaluation

of data from actual test administrations should examine how well the test continues to meet the needs of all test takers.

Example. Assembling items or tasks into a test involves many factors such as the length of a test, the way items are laid out on the page, whether the test is computer-administered, and so on, that require input and review from diverse groups similar to those used for reviewing individual items. Once tests have been assembled, they should be field-tested with trial samples that include members of the all disability categories that the test is intended for, even if those sample sizes are too small for some statistical analyses. Whenever possible, large enough samples should be used for the field trial so that comparisons of the item and test characteristics, as well as comparisons of the internal structure of the assessment for the various student groups, can be carried out and used to investigate score comparability for these groups.

Guideline 3-D. Document the steps that have been taken to ensure that the characteristics of all test takers have been considered.

The steps that have been taken to ensure accessibility should be documented and communicated so that assessment professionals can examine evidence to evaluate whether required test development and implementation steps have been followed and so that end users can evaluate how well a test will meet their needs and what interpretations are appropriate or not appropriate to make.

Example. Provide professional evaluators and test users with materials that document which test takers' needs have been taken into account throughout the test design, development, and implementation processes. The tests' technical materials and manuals should include discussion of the rationale for any recommended accommodations and the evidence that the accommodations support more valid testing.

Principle 4: Reading assessments reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student’s proficiencies.

Rationale. Although a goal in designing and developing reading assessments that are accessible is to minimize the need for accommodations, some students may still need them. To be accessible, reading assessments first must attempt to build accessibility throughout the test itself, often by integrating tools that have functions similar to some accommodations. When this cannot be accomplished, accessibility considerations require that the assessment allows a variety of accommodations that address students’ needs, and produce valid results. The accommodation should allow students to show what they know and can do in grade-level reading rather than merely show the effect of their disabilities. The assessment indeed must be reviewed to make sure all access tools that can be integrated into the assessment have been integrated. Next, it is necessary to determine whether other accommodations are needed. If they are, the technical adequacy of the assessment must be addressed and information and support supplied to ensure that the accommodations are provided in an appropriate manner. There may be other adjustments to the assessment that are needed for those instances where the student’s disability precludes the performance of a skill that is required by the reading assessment, for example, items that test homophones (words with the same pronunciation but different meaning and spellings) for students who are deaf. All of these steps should be undertaken without compromising the comparability and validity of inferences that are made about students’ reading knowledge and skills.

Guideline 4-A. Begin the assessment development or revision process by reviewing allowed accommodations to determine whether they could be incorporated into the design of the assessment.

Questions about accommodations and design of the assessment should be asked each time there is an assessment redesign, a change in accommodation policies, or an adjustment to the content standards.

Example. If a state redesigns its reading assessment for a computer-based platform, the state should re-examine whether any of its allowed accommodations could be incorporated into the assessment rather than provided as add-ons. In this situation, accommodations that could be incorporated include highlighting text via computer technology or offering screen readers. When using screen readers, care should be taken so the text of the assessment meets criteria for easy access by screen readers. Similarly, descriptions of pictures in the assessment should be provided to ensure compliance with Section 508 requirements and full usability for the intended purpose of the picture. An untimed test is a good example of something previously considered to be an accommodation (extra time), but now is generally accepted as a characteristic of the assessment.

Guideline 4-B. Identify and determine the essential accommodations that are still

needed after incorporating as many as possible into the assessment.

Despite best efforts to account for the needs of all test takers throughout the various phases of test development and evaluation, accommodations might still be needed. The purpose of the accommodation would be to make tests fairer and improve the inferences that can be made from some students' test scores without altering the focal construct.

Example. A paper-and-pencil version of a computerized assessment might be needed for students who require magnification too high to be feasible for scrolling and other navigation requirements. Some students might need separate settings or multiple-day testing, both of which are considered to be accommodations in many states. Unique accommodations may also be needed, such as providing prompts to slow down students who are not paying full attention to the task and are moving through the assessment too quickly.

Guideline 4-C. Develop a strong rationale and evidence to support the validity of inferences from assessment results when accommodations are provided.

For accommodations that cannot be incorporated into the assessment itself, a strong rationale needs to be developed about the appropriateness of the accommodations and reasons should be provided for why results are valid when the accommodations are used. Evidence needs to be included about students' experiences with the accommodations during instruction as well. Whenever possible, the comparability and validity of inferences based on scores from accommodated assessments should be researched and evaluated.

Example. Strong rationales are those that clarify the intended content to be measured by the assessment and relate this to the disability characteristics of those students needing the accommodation. The rationales also document that the needed accommodations are regularly used in instruction to support the acquisition of the content being assessed and that the accommodations have been included during field testing. Empirical studies of construct equivalence, such as differential item functioning and factor analysis studies, may be cited or undertaken to provide evidence to support the comparability and validity of inferences based on scores from accommodated assessments.

Guideline 4-D. Provide information and support to ensure that accommodations are implemented in an appropriate manner.

The validity of results from assessments taken with accommodations rests not just on the relation of the accommodation to the construct being measured, but also the appropriate implementation of the accommodations. It is the responsibility of the state or the test developer to ensure that information and other support are provided to decision makers and students so that accommodations are implemented appropriately. Evidence of appropriateness includes students using accommodations designated for them, participation by students in the decision-making process, and the provision of

accommodations running smoothly without disturbing other students.

Example. One aspect of providing information and support for accommodations is to ensure that there are training materials for individualized education program (IEP) teams so that their decisions about accommodations are appropriate for the assessment. Another aspect is to provide guides that assist in the appropriate administration of an accommodation, such as the read-aloud accommodation. If this accommodation is not embedded in the assessment, it is subject to possible administration errors if specific guidelines about implementation are not available.

Implementing accommodations in an appropriate manner also means ensuring that students actually use accommodations designated for them. To achieve this, states and test producers should ensure information is available for schools to use to educate all students about the purpose and acceptability of accommodations, and to inform the students who use accommodations about the importance of the accommodations in obtaining accurate measures of their knowledge and skills. Finally, support should be in place to engage students in the process of selecting accommodations, so that those accommodations provided are ones that the student will use.

Guideline 4-E. Adjust the reading assessment approach to address the needs of some groups of students that cannot be met by typical test design or accommodation procedures.

In a few cases, assessments may contain items derived from state standards where a student's disability precludes the performance of the skill. For example, items that test homophones are considered inappropriate for students who are deaf or hard of hearing. A systematic way to address these situations is needed, which might involve making an adjustment to the assessment.

Example. A process to examine the kinds of adjustments that might be acceptable includes checking what the test is intended to measure, determining whether additional accommodations might be provided, and considering dropping items or identifying replacement skills. Specifically, states would start by reviewing whether the content standard reflected in an item needs to be assessed if it cannot be assessed appropriately for all students. Given the decision that the standard should be assessed, the state may want to determine whether there are accommodations that could address students' needs. This should be considered before assembling the test so that special forms, for example, can be created as necessary (such as a braille form).

If no accommodations are appropriate, there are several approaches that can be taken. One simple approach might be to delete a few questions from the test that are inaccessible for some student groups in the population. This approach might be possible if the reduced test still provides a comparable and reliable score that covers the state standards. Another approach might be for the state to consider identifying replacement skills for the subset of students for whom the standard is inappropriate. Depending on

the replacement skills identified, the adjusted test may or may not provide scores that can be considered comparable to the unadjusted test scores.

Whenever the assessment is adjusted, it is important to determine if scores on the original and adjusted assessment can be considered to measure the same underlying construct. If there is evidence that this is not the case, those using scores from the assessment should be informed that the adjustment to the assessment may have changed the construct measured by the assessment.

Principle 5: Reporting of reading assessment results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results.

Rationale. A great deal of effort goes into the creation of scores for any large scale testing program. Care is taken to assure that the test is designed and developed with a high level of quality and that attention is paid to meeting professional standards for the assessment. Focus is often on the reliability, validity, and fairness of the assessment, as well as other characteristics of the test scores. Because it is the testing program score reports that convey the final results of the assessment to the test taker, teacher, parent, policymaker, and so on, it is important that the same level of care be invested in score reporting as is in other aspects of the testing program. Assuring that the score reports that are produced by a testing program support the correct interpretation and use of the scores is of primary importance to the testing program and to all score recipients. It is particularly important that testing programs consider the needs of a diverse population of test takers and parents, such as students with disabilities and English-language learners, when designing score reports and score reporting procedures.

Guideline 5-A. Provide clear and concise score reports that are appropriate for a diverse audience.

A requirement of the No Child Left Behind (NCLB) legislation is that student results be reported in a clear and easy to understand manner. There are a number of ways to increase the understandability of score reports. For example, decreasing the statistical jargon will be helpful to everyone, but may be particularly helpful to English-language learners. Reporting results in multiple ways, such as numerically and graphically, may provide more understandable results to a more diverse group of score recipients. In addition, when possible, score reports should be provided in electronic formats that are accessible using assistive technologies such as screen readers. Practices such as highlighting sections and headings are another way to increase the understandability of score reports. Also, assessment results will be interpreted with greater understanding if the reports contain clear statements of the purpose of the assessment, an explanation of how the results should and should not be used, a description of the scores, and some measure of the precision of the scores that is described clearly and succinctly. Finally, all score reports from large scale assessments should contain information about how students did on the test, what kind of skills or knowledge the test performance reflects, and what can be done to help the child improve.

Example. A well-designed score report should provide answers to questions that are important for the score report recipients. Students will want to know what their scores are; the score should be displayed prominently and highlighted so it is easy to find. Parents will want to know how their children's scores compare to other scores. When information such as a child's percent correct on a particular set of items is presented, it should be explained in language that the score report user can understand.

Guideline 5-B. Pilot score reports with all relevant groups of score users.

Score reports should be carefully piloted before they are actually used to report scores. These pilots can take several forms; for example, surveys or focus groups can both be very effective. The key to the success of the pilot studies is the inclusion of all relevant groups in the pilot. Teachers may have very different requirements for, and opinions about, what is included in a score report than students or school administrators. A parent of a child with a disability may have a very different need for the development of a particular component of the score report than a teacher has in mind. The parent of an English-language learner may find the language of the score report particularly confusing. Consequently, it is essential that all stakeholders in the assessment be included in the score report pilots.

Example. An effective way of piloting score reports is to present a mocked-up score report to a focus group of potential score users and ask specific questions about different sections of the score report. Each focus group should contain a cross-section of all potential recipients of the score reports. In addition, the focus group sessions should be held in different locations (e.g., urban, rural) to ensure a wide variety of input to the design of the score reports.

Guideline 5-C. Detailed information about the assessment and score results is available in a document that is accessible to all test takers and score users.

Publishers following good testing practice provide information that helps score users understand score reports in an interpretive guide or technical manual. The guide or manual should include all the information that is necessary to understand the purpose and meaning of the test scores and to interpret the test score results. Comparative information on group performance may be included in the guide. Studies that have been carried out to evaluate the properties of the test scores (e.g., reliability, validity) should be included. When interpretive guides and technical manuals are prepared, it is very important to consider all groups of students that will be tested. For example, testing programs should make an effort to report the reliability of test scores for diverse populations such as students with disabilities and English-language learners whenever sample sizes permit this type of analysis. In addition, the interpretive guide should include a discussion of test modifications or accommodations used by students with disabilities and English-language learners and how to interpret scores obtained under these nonstandard conditions.

Example. An interpretive guide could be presented in the form of a folder that accompanies the score report. The folder could have interpretive information printed on the inside and back pages and it could contain a pocket with a sealed envelope for the student's score report. The interpretive guide should contain only information that the score recipient can use, and any complex information such as a standard error or a *p*-value must be explained in language the score recipient will understand. An explanation of how the information can be of use to the score recipient should also be included.

Guideline 5-D. Provide information regarding the precision of reported scores for all relevant groups.

It is important to gather evidence of the reliability, validity, and fairness of the scores for all relevant groups of the target population, including students with disabilities who have taken the test with modifications or accommodations or who have taken the test under standard conditions. Whenever sample sizes permit, reliability, fairness, and validity studies should be carried out on the scores obtained by students with disabilities and English-language learners on standard and nonstandard administrations of tests of reading achievement. The information obtained in these studies should be documented and made available to recipients of score reports.

Because sample sizes vary from state to state and disability to disability, it may not be possible to provide statistical information for all possible subgroups taking the test. However, there are some subgroups of test takers, such as students with learning disabilities, that are typically large enough to support most reliability and differential item functioning analyses. In addition, it is sometimes possible to accumulate data over several test administrations, or over several years, such that sample sizes accrue, making it possible to carry out statistical analyses on some of the less prevalent disability groups and smaller accommodation groups.

Example. Large scale testing programs that administer standardized reading assessments for accountability purposes typically report statistics such as reliability, standard errors of measurement, average item difficulty level, and differential item functioning statistics for males and females, ethnic groups, and sometimes English-language learners. This type of reporting is sometimes possible for test takers with disabilities such as learning disabilities. When these statistics are reported, they should be accompanied by appropriate explanations so that a lay person can understand the implications of the information.

Conclusions

The principles and guidelines in this document serve not only as a vision for future assessments but also as a road map for improving the accessibility of current assessments. They were developed in response to the many challenges that exist in obtaining valid measures of the reading knowledge and skills of students with disabilities.

Recall that the goal was not to devise ways to hold students with disabilities to different performance standards, but rather, to address the many ways in which standards-based reading assessments can become more accessible for all students and in doing so, ensure that they measure the same knowledge and skills at the same level as traditional large scale reading assessments. The principles and guidelines in this document do this by addressing critical aspects of reading assessments, including the population of students who must be accurately measured, the nature of the reading content assessed, the development process, the role of accommodations, and the reporting of results.

Although the principles are to be viewed as a whole, it is recognized that some guidelines, such as those focused on technology, may not be possible to pursue immediately. They should, however, be in the vision for future assessments, and the essence of the principles and guidelines adhered to even though technology, for example, may not yet be used for assessment.

These guidelines likely will evolve as knowledge about assessing students with disabilities increases. The guidelines can be strengthened and clarified as relevant research is conducted and other experiences and evidence are compiled. The National Accessible Reading Assessment Projects (NARAP) will continue to do this as long as funding permits.

The summary of the current support for the principles and guidelines in the appendix is viewed as evolving as well. The hope is that the annotated bibliography will become a living document, with annotations added as appropriate. If necessary, principles and guidelines could be revised over time to reflect the new research, experience, and other evidence.

Although these principles and guidelines were written for large scale reading assessments, they are applicable to other types of assessments and other content areas. Similarly the principles and guidelines were directed to state assessment personnel and test developers, but can be used by all individuals who create and use assessments.

Appendix

Support for Guidelines

The accessibility principles for reading assessments are supported by research, existing standards, and the consensus of experts. These supports range from highly rigorous experimental research to theoretical treatises on relevant topics. The purpose of this appendix is to provide users of the principles with an annotated list of support for each guideline included in each of the five principles. The list is not exhaustive in order to keep the appendix to a reasonable length; however, it is a fair representation of the research and other reports available for review.

This appendix is a living document. In other words, new research, standards, and evidence will be added as they emerge in support of the guidelines that underlie the principles. Readers are invited to share either confirmatory or contradictory support with NARAP researchers to ensure that the supporting information for the accessibility principles for reading assessments is complete.

Principle 1: Reading assessments are accessible to all students in the testing population, including students with disabilities.

Guideline 1-A. Understand and account for the range of student characteristics and experiences that impact reading when designing reading assessments.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

The *Standards for Educational and Psychological Testing* were developed to support the “sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices” (p. 1). Standard 10.5 addressed the need to ensure that inferences from tests reflect the intended construct rather than students’ disabilities. This standard stated, “In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement” (p. 106).

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

This book demonstrated how student characteristics can be considered as a part of making decisions about the nature of assessments that are used for tracking, promotion, and graduation of students. The book included chapters on students with disabilities and English language learners.

Johnstone, C. J., Bottsford-Miller, N., & Thompson, S. J. (2006). *Using the think*

aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners (Tech. Rep. No. 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

This research was designed to examine the appropriateness of assessment items for individual students. A think-aloud process was used, with students talking about each item, so that researchers could determine how students approached multiple-choice and constructed-response items. Researchers found that the process revealed differences among students who had different characteristics, allowing the researchers to determine how students worked through each item and where stumbling blocks in the design were for students.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

This document is a guide for professionals “to provide and use tests that are fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics” (p. 2). Originally developed in 1988 to ensure fair testing for all test takers, the *Code of Fair Testing Practices in Education* was recently updated and aligned with the *Standards for Educational and Psychological Testing* (1999). Principles for both test developers and test users were provided. Guideline A-9 in the *Code of Fair Testing Practices in Education* indicated that test developers should “obtain and provide evidence on the performance of test takers of diverse subgroups” (p. 4) and that evidence should be evaluated to ensure that differences in performance are related to the skills being assessed. This implies that the differences should not be due to disabilities alone and that there is a need to provide evidence of this.

Koenig, J. A., & Bachman, L. F. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*. Washington, DC: National Academies Press.

This report from the Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large Scale Assessments addressed critical issues in the assessment of English language learners and students with disabilities, especially for the National Assessment of Educational Progress (NAEP), but relevant to other assessments as well. Recommendation 4.3 was: “NAEP officials should more clearly define the characteristics of the population of students to whom results are intended to generalize. This definition should serve as a guide for decision making and the formulation of regulations regarding inclusion, exclusion, and reporting” (p. 5).

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127–143.

This article explored factors affecting score differences on computerized and paper-and-pencil reading tests. The authors found that score differences among 2,000 students across four grades were less related to item characteristics and more affected by such student characteristics as response style variables (e.g., omissions) and socioeconomic status (determined by a student's free lunch eligibility), although response style and socioeconomic status were relatively independent of each other. These findings highlight the importance of attending to student characteristics and experiences when designing reading assessments.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Rep. No. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

This report presented a set of elements essential in assessments that are designed to be appropriate for the widest range of students without changing the construct that the assessment is intended to measure. Seven elements were proposed, the first of which was inclusive assessment population. The report argued that assessments need to be responsive to increased diversity, increased inclusion of all types of students in the general curriculum, and increased emphasis and commitment to serve and be accountable for *all* students.

Thurlow, M. L., Quenemoen, R. F., Lazarus, S. S., Moen, R. E., Johnstone, C. J., Liu, K. K., Christensen, L. L., Albus, D. A., & Altman, J. (2008). *A principled approach to accountability assessments for students with disabilities* (Synthesis Rep. No. 70). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

In a report on qualities of inclusive systems of assessment that are used for accountability, a central principle is "All students are included in ways that hold schools accountable for their learning" (p. 4). A characteristic under this principle says, "The validity of the system is assured through technically defensible assessments that address the implications of varied student learning characteristics," which requires "careful consideration of varied student learning characteristics" to ensure that "assessment design processes build on understanding how all students learn and show what they know" (p. 6).

Winograd, P., Flores-Duenas, L., & Arrington, H. (2003). Best practices in literacy assessment. In L. M. Morrow, L. B. Gambrell, & M. Pressley (Eds.), *Best practices in literacy instruction* (2nd ed., pp. 201-238). New York: Guilford.

This chapter provided a list of 18 best practices in literacy assessment drawn from seven sources published in the mid- to late-1990s. One best practice particularly relevant to Guideline 1-A encouraged the development of assessments that focus on the full range of student characteristics: "Provide educators and others with richer and fairer information about all children, including those who come from linguistically and culturally diverse backgrounds" (p. 209).

Guideline 1-B. Begin the development of reading assessments by applying the elements of universal design.

Allan, J., M., Bulla, N., & Goodman, S. A. (2003). *Test access: Guidelines for computer-administered testing*. Louisville, KY: American Printing House for the Blind.
Available at: <http://www.aph.org/tests/access/access.pdf>

This publication addressed issues that may arise for test takers who have disabilities, particularly those with low vision or who are blind. In spite of advances in technology (e.g., computer-administered tests), there is still an achievement gap between students with visual impairments and those without. The authors presented the principles of inclusive design and argued that tests must be made accessible to all potential test takers, regardless of test format or test-taker disability. Theoretical arguments and practical solutions were recommended for designing tests that are accessible from the beginning, rather than depending on accommodations for access. The authors also stressed that students must be tested in the format in which they typically learn (including allowances for assistive technology use).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

The *Standards for Educational and Psychological Testing* support inclusiveness by recommending that “all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure” (p. 74). Likewise, the standards addressed accessible test items by recommending that test developers “research whether any items are more difficult for students from particular subpopulations. This can be accomplished through the administration of a field-test that can help determine item difficulty and ability to discriminate among test takers of different standing on the scale” (p. 39).

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

This article reviewed three approaches that can add validity evidence to states’ item-review processes. The first is a structured sensitivity review process that focuses on universal design considerations for items. The second is a series of statistical analyses intended to increase the limited amount of information that can be derived from analyses on low-incidence populations (such as students who are blind, deaf, or have cognitive disabilities). Finally, think-aloud methods are described as an approach for understanding why particular items might be problematic for students.

Salend, S. (2009). Using technology to create and administer accessible tests. *Teaching Exceptional Children*, 41(3), 40-51.

This article provides an overview of how the principles of universal design can be incorporated into testing via the use of technology. The article summarizes the principles of universal design and provides a link to testing and an example of how the principle can be implemented. The author provides concrete examples of how to use technology to select typographic and visual elements, as well as how technology can enhance motivation through student choice, feedback, and adaptive testing. While some of the article is specifically focused on classroom assessments, the author provides an excellent overview of the assessment needs of students and the types of technology that can be used to improve testing for students with a variety of characteristics.

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zieky described several steps for conducting item reviews on licensure tests. Many of his suggestions are transferable to K-12 large-scale assessment, including several that mirror the principles of universal design of assessment: (a) establish an advisory group specifically for fairness issues; (b) include only content in the test that is clearly justifiable; (c) provide guidelines about avoiding sexist, racist, stereotyped, or offensive language, illustrations, or stimulus materials to reviewers and item writers; (d) ensure that linguistic demands are concordant with the test's purpose; (e) treat test takers equally regardless of personal characteristics that are not relevant to the test; and (f) provide accommodations for people with disabilities (or others as allowable). According to the author, a structured and documented approach to reviewing items may increase the quality and fairness for the assessment population.

Guideline 1-C. Use technologies and other evidence-based approaches to provide all students with a variety of assessment options within a similar testing experience.

Dolan, R. P., Burling, K. S., Harms, M., Beck, R., Hanna, E., Jude, J., Murray, E. A., Rose, D. H., & Way, W. (2009). *Universal design for computer-based testing guidelines*. Iowa City, IA: Pearson.

In these guidelines the authors use the principles of Universal Design for Learning (UDL) as a foundation for development of computer-based testing platforms and items. These guidelines are organized according to three tiers: (a) test delivery considerations, (b) item content and delivery considerations, and (c) component content and delivery considerations. The component (e.g., text, audio, video, interactive manipulatives) content and delivery considerations tier is further organized according to the various categories of processing that students apply during testing, based on a framework derived from UDL principles. The resulting guidelines are designed to facilitate creation of items that maximize the benefits of technology and minimize the impact of construct irrelevant variance.

Edyburn, D. L. (2007) Technology-enhanced reading performance: Defining a research

agenda. *Reading Research Quarterly*, 42(1), 146-152.

This article discussed the use of technology to assist readers with cognitive disabilities. The author identified four categories of issues that he considered fundamental to understanding the efficacy of technology for enhancing reading performance (what does it mean when we say a person is a reader or nonreader, instructional methods, remediation versus compensation, measurement of outcomes) and provided new directions in each category. The author argued that we should not assume that academic performance achieved without the aid of external devices and resources is necessarily to be valued more than performance that is dependent on tools or resources. Recognizing that the topic is controversial, the author nonetheless called for “significant ethical, theoretical, and empirical work regarding the nature of assistive technology for enhancing reading performance” (p. 151).

Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice*, 16, 174-181.

This article introduced a data-based approach as a way to help teachers make decisions about testing accommodations for individual students with learning disabilities. In the process of suggesting that a data-based approach is needed for selecting accommodations, the authors described difficulties with other approaches for selecting accommodations. They discussed challenges in basing judgments about which accommodations to use on the existing research literature, pointed out the difficulty of applying a single accommodation to a group as heterogeneous as students with learning disabilities, and cited research on problems with using a teacher’s judgment to select an accommodation for any given student. Such challenges encourage the adoption of accessible assessment solutions that do not require assigning different accommodations to different students. This implies that using approaches or designs that incorporate options that are the same for all students is preferred.

Ketterlin-Geller, L. (2005). Knowing what all students know: Procedures for developing universally designed assessments. *The Journal of Technology, Learning, and Assessment*, 4(2). Available from <http://www.jtla.org>

This article explored the use of technology as a way of improving accessibility of large-scale assessments. The author described the use of a computerized testing system for on-demand accommodations based on student need as determined from results of a pre-test. The underlying theoretical argument was that assessment formats that allow students to access accommodations, such as changes in font, delivery, and language load, on demand provide a more universal approach to assessing all students.

Minnema, J., Thurlow, M., & Hopfengardner Warren, S. (2004). *Understanding out-of-level testing in local schools: A first case study of policy implementation and effects* (NCEO Out-of-Level Testing Project Rep. No. 11). Minneapolis, MN: University of

Minnesota, National Center on Educational Outcomes.

In this case study, interview data were collected from educators and students to learn more about the experience of using out-of-level testing. Teachers noted inappropriate behavior by some students taking out-of-level tests in a classroom where everyone could see that some students were taking different tests. The teachers concluded that the inappropriate behavior resulted because some students were upset about taking tests at a lower level than other students. This illustrates how students can respond negatively to test experiences that are supposed to be for their benefit if those experiences mark them as different or embarrass or stigmatize them. Accessible assessment that permits all students to have the same experience avoids stigmatizing students by treating them differently, thus suggesting that the use of options that are the same for all students, rather than targeted to one or a few, are preferable.

Salend, S. (2009). Using technology to create and administer accessible tests. *Teaching Exceptional Children, 41*(3), 40-51.

This article provides an overview of how the principles of universal design can be incorporated into testing via the use of technology. The article summarizes the principles of universal design and provides a link to testing and an example of how the principle can be implemented. The author provides concrete examples of how to use technology to select typographic and visual elements, as well as how technology can enhance motivation through student choice, feedback, and adaptive testing. While some of the article is specifically focused on classroom assessments the author provides an excellent overview of the assessment needs of students and the types of technology that can be used to improve testing for students with a variety of characteristics.

Guideline 1-D. Document decisions that are made to make tests more accessible, and monitor the effects for different groups of students.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Standard 7.9 of the standards developed to support good testing practices addressed the need to examine the likely consequences of test use. This standard stated, “When tests or assessments are proposed for use as instruments of social, educational, or public policy, the test developers or users proposing the test should fully and accurately inform policymakers of the characteristics of the tests as well as any relevant and credible information that may be available concerning the likely consequences of test use” (p. 83).

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Guideline C-6 in the *Code of Fair Testing Practices in Education* indicated that test developers should “provide information to enable test users to accurately interpret and

report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results” (p. 8). This guideline noted the importance of documenting what was done to include or exclude students and the possible effects of the decisions on the results.

National Academies. (2007). *Lessons learned about testing: Ten years of work at the National Research Council*. Washington, DC: Author.

This booklet addressed the findings of National Research Council panels. One of the lessons identified in the booklet was, “New testing programs should build in an evaluation component” (p. 13). Specific reference was made to the need to study the impact of a testing program on particular groups of students. This concept is critical to the notion of monitoring the effects of assessments, including those assessments where decisions are made to increase accessibility.

Principle 2: Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence, and attuned to accessibility concerns.

Guideline 2-A. State standards are grounded in an evidence-based definition of reading.

Afflerbach, P. (2004). *National Reading Conference policy brief: High stakes testing and reading assessment*. Oak Creek, WI: National Reading Conference.

This policy brief, one of several commissioned by the National Reading Conference to review the empirical research on particular topics for use by the general public, stated that assessments “must be clearly and carefully tied to an informed understanding of what reading ‘is’” (p. 3). “Assessments should provide clear distinctions between the acquisition of reading skills and the effective use of the skills for various purposes. We have voluminous evidence that reading is developmental in nature. . . . An assessment that describes students’ ability to decode single syllable, phonetically regular words is important for early readers, as is assessment that describes more advanced readers’ ability to critically interpret and evaluate persuasive writing” (p. 13).

International Reading Association. (1999). *High stakes assessments in reading: A position statement of the International Reading Association*. Newark, DE: Author.

The International Reading Association developed a policy statement that provided “a call for the evaluation of the impact of current types and levels of testing on teaching quality, student motivation, educational policy making, and the public’s perception of the quality of schooling” (p. 1). One key recommendation of the brief was to “design an assessment plan that is considerate of the complexity of reading, learning to read, and the teaching of reading” (p. 9).

Johnston, P., & Costello, P. (2005). Principles for literacy assessment. *Reading Research Quarterly*, 40(2), 256-267.

In this review, Johnston and Costello argued that a focus on the *what* of assessment is a disservice to society. They asserted that attention must be paid to the *how* of assessment. The authors cited research that more completely explicates the idea of literacy. The authors noted that literacy is more than a set of skills and strategies. It also involves *identities, values, and dispositions* such as the resilience and reciprocity necessary for a democratic society. Clarification of the *how* of literacy assessment is vital because what educators assess is often what gets taught. Literacy assessments, especially ones that are high stakes, must be designed to teach the complete range of literacy's dynamics. The authors contended that literacy assessments must reflect the fact that "different tools and social contexts invoke different strategies and ways of thinking" (p. 262). They offered evidence that authentic assessments and formative assessments should occupy a central role in educational practice and demonstrate the efficacy of such assessments. While cognizant of the embedded nature of the current, narrowly focused literacy assessment in teachers' pedagogy and in society at large, the authors maintained that when a more complete understanding of literacy informs socially situated assessment, just relationships will be promulgated and democratic education will be achieved.

McKenna, M. C., & Stahl, S. A. (2003). *Assessment for reading instruction*. New York: Guilford Press.

Stating that all reading assessment is based on a model, the authors provided support for the contextual model, based on research and a definition of reading that outlines how children read. The contextual model views individual students as having attitudinal, motivational, or cognitive needs that instruction should address and that assessment must take into account. The authors concluded that a model that grounds reading assessment is key for making sense of an individual student's reading assessment data and how this information should inform instruction. This model also guides the selection of the best measures to inform educators about students' needs.

Moje, E. B., Dillon, D. R., & O'Brien, D. G. (2000). Re-examining roles of learner, text, and context in secondary literacy. *Journal of Educational Research*, 93(3), 165-180.

As researchers have expanded concepts of reading to include social and cultural perspectives, the authors of this article argued that the role of the reader, text, and context have also evolved. The authors indicated that definitions of reading need to consider readers as having a more central role in the process of reading; texts as being diverse, intertextual, and constantly changing; and contexts as being critical to shaping meaning. The authors concluded that reading researchers need to broaden definitions of reading to include the reader, the text, and the context because all are important to how we define literacy and how we measure students' abilities.

Paris, S. G., & Hoffman, J. V. (2004). Reading assessments in kindergarten through third grade: Findings from the Center for Improvement of Early Reading Achievement. *The Elementary School Journal*, 105(2), 199-217.

This article synthesizes several studies on early reading assessment designed to determine the kinds of assessments available to teachers and the teachers' reactions to the assessments (e.g., teachers' use of informal reading inventories for formative and summative purposes; innovative assessments of children's early reading). The authors draw several conclusions: (a) researchers must continue to investigate the ways in which assessment tools can be broadened to focus on multiple factors and the interaction of these factors in ways that reflect authentic learning and teaching environments; (b) the gulf between what teachers value as informal assessments and what is imposed on them in the form of standardized testing appears to be broadening; and (c) researchers cannot lose sight of the fact that good assessment rests on good theories of reading, teaching, and learning.

Pearson, P. D., Barr, R., Kamil, M. L., & Mosenthal, P. (Eds.). (1984). *Handbook of reading research* (Vol. I). Mahwah, NJ: Lawrence Erlbaum Associates.

Barr, R., Kamil, M. L., Mosenthal, P., & Pearson, P. D. (Eds.). (1996). *Handbook of reading research* (Vol. II). Mahwah, NJ: Lawrence Erlbaum Associates.

Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (Eds.). (2000). *Handbook of reading research* (Vol. III). Mahwah, NJ: Lawrence Erlbaum Associates.

Several chapters in the three volumes of the *Handbook of Reading Research* reviewed various models of reading and noted what researchers have learned from these models that impacts reading research, instruction, and assessment. A number of models focused on word recognition, such as Gough's work and LaBerge and Samuels' early ideas on a linear perspective; Rummelhart's research and Stanovich's early work on an interactive processing perspective; and work that focused on comprehension processes by Just and Carpenter and Kintsch and by Van Dijk. Overall conclusions indicated that not all models build on previous models, but using features of several models allows reading researchers the ability to explain a great deal about the reading process. However, early reading models have gaps, such as in the areas of the role of schemata and metacognition. Current models have indicated that reading is more than just a cognitive process—it is a developmental process as well. Bloome and Green's research also noted that reading is a social and linguistic process.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, 16, 32-71.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford.

These two publications described models of reading, based on experimental research. The author referred to Rummelhart's interactive model of reading, which specifies that proficient readers efficiently store and activate various schemata when needed either in a top-down or bottom-up fashion. While other reading researchers have alluded to the potential of an interactive processes model, few have discussed the relationship of this model to the nature of individual differences in reading fluency. In these two works, Stanovich introduced the now classic interactive compensatory model of reading to address problems with either top-down or bottom-up processing so that a deficit in any one process will result in greater dependence on other knowledge sources, regardless of their level in the processing hierarchy (whether they be more at the "top" or the "bottom"). Stanovich noted that proficient readers do not simply use the redundancy "inherent in natural language to speed word recognition," instead, they engage in "general comprehension strategies and rapid context-free word recognition" (p. 32).

Guideline 2-B. Design reading tests to allow for individual measurement of the components reflected in state standards.

Afflerbach, P. (2004). *National Reading Conference policy brief: High stakes testing and reading assessment*. Oak Creek, WI: National Reading Conference.

The purpose of this policy brief was to guide the design of large scale reading assessments. The document called for the measurement of reading skills beyond the standardized multiple-choice comprehension assessments and the provision of useful feedback to students and teachers. It also recommended that assessments should provide students with useful information about their developmental accomplishments with clear suggestions for improvement and teachers with useful diagnostic information that can be linked to classroom instruction.

Buly, M. R., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis*, 24(3), 219-239.

This study probed beneath students' failing scores on a state reading assessment to investigate the needs of struggling students and implications for assessment policy. To deconstruct students' poor reading performance, Buly and Valencia used multiple measures that assessed word identification, phonemic awareness, comprehension, reading fluency, and vocabulary. They found that drastically different instructional approaches were needed to accommodate various students' needs, and they considered the implications of these findings for ongoing state and local educational reforms. The authors concluded that "simple percentages of students failing the test masked empirically derived components of reading ability: meaning (comprehension and vocabulary), fluency (rate and expression), and word identification. Furthermore, even average group scores in each of these components did not tell the real story. Instead, we found that students exhibited several distinctive patterns of performance that contributed to their poor showing on the state reading assessment. Reading failure is multifaceted and it is individual" (p. 232).

Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., & Vaughn, S. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children*, 72, 136-150.

This research study is an example of how a state department of education collaborated with researchers to develop an accommodation policy that was in keeping with the construct to be assessed and to design a research study to test the merits of the accommodation on test scores for two populations of students. The authors of this article evaluated accommodations that were specifically designed for students with word-decoding problems. Due to different components of reading comprehension, these students may comprehend text adequately but their word recognition skills and fluency were weak. The authors sought to decrease the impact of word recognition on high stakes tests by increasing comprehension through accommodations (based on an interaction hypothesis: Adding these accommodations would not improve scores for students with adequate decoding skills). The set of accommodations that were used included reading comprehension stems, proper nouns, and possible answers. Of the 187 third grade students who participated, half had dyslexia. Of the students with dyslexia, 44 students were given accommodations and 47 were not; of the students with adequate fluency skills, 44 were given accommodations and 47 were not. The high stakes test used in this study was the Texas Assessment of Knowledge and Skills. A mixed-model analysis of covariance was used to analyze the data. There was a statistically significant increase in the performance of poor decoders who had accommodations and an increase in the number of students who passed the test.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities from identification to intervention*. New York: Guilford Press.

This book illustrated the importance of developing assessments that allow students to demonstrate their comprehension skills independently of their word recognition and reading fluency skills. The authors summarized research evidence to support three subtypes of reading-based learning disabilities: primary deficits in (a) word recognition, (b) fluency, and (c) comprehension, with the strongest support for word recognition deficits (e.g., dyslexia). This evidence—that a pattern of skill deficits isolated to one component of reading is observed in students with learning disabilities—supports the assessment of each component in isolation so that a deficit in one component of reading does not preclude showing proficiency in another component.

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

The use of multistage tests to assess skills and abilities for students with disabilities has increased recently. It is possible to use multistage testing to provide a better match of a test's difficulty and content to a student's skills or abilities and consequently to obtain a better estimate of the student's proficiency level. In a two-stage test, a student is given a relatively short test, and based on the score, he or she is routed to one of several second

tests. The second test measures grade-level standards. “This article describes multistage tests, including two-stage and testlet-based tests, and discusses the relative advantages and disadvantages of multistage testing as well as considerations and steps in creating such tests” (p. 1).

Johnstone, C. J., Thurlow, M. L., Thompson, S. J., & Clapper, A. T. (2008). The potential for multi-modal approaches to reading for students with disabilities as found in state reading standards. *Journal of Disability Policy Studies, 18*, 219-229.

This article reviewed state reading standards from 49 states. Broad themes that were drawn from states’ standards documents showed standards ranging from specific skills (e.g., decoding, fluency, and phonemic awareness) to standards that addressed students’ ability to interpret, mine information, and connect with literature for personal reflection and growth. Only the foundational skills are ones that need to be accessed through visual or tactile modalities, suggesting that a different approach (e.g., auditory) might be used for the other skills. This kind of differentiation in possible approaches suggests that assessments could be designed to measure different standards with different methods, a strategy that would support allowing for individual measurement of the components reflected in state standards.

Laitusis, C. C., & Cook, L. L. (2008.) Reading aloud as an accommodation for a test of reading comprehension. *Research Spotlight, 1*, 15-20.

This article summarized a series of research studies designed to examine the comparability of reading comprehension test scores from standard and audio presentation (read-aloud accommodation) testing conditions. Studies included repeated measures analysis of variance with an experimentally designed random assignment data collection, differential item functioning, and factor analyses. Results indicate that it is possible to report comprehension test scores (from audio presentation and standard administration) on the same scale when foundational skills are not part of the construct being assessed. However, the authors suggested including a measure of fluency in the assessment when audio presentation is included as an accommodation.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185–187.

Multistage tests provide a flexible way to better target the ability levels of a heterogeneous group of examinees. Test takers are given a short test, and based on how they score on this test, they are branched, or routed, to a test that is the best match to their ability level. The second test is typically selected to measure grade level standards. Multistage tests can be multitiered and contain three or more stages. This article provides a brief history of multistage testing and contrasts this type of testing with computer adaptive assessment. The author discusses some of the practical issues associated with this type of testing.

Paris, S. G., & Stahl, S. A. (Eds.). (2005). *Children's reading comprehension and assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

The overview statement of the “Historical and Theoretical Foundations” section in this book indicated that several chapters discussed the idea of subskills in regard to comprehension. Pearson and Hamm’s chapter stated that there are several identifiable subskills or components of the comprehension process. Kintsch and Kintsch’s chapter discussed subskills in extracting meaning from text, a situation model, and reader’s prior knowledge. Sweet’s chapter argued that there are operations in the comprehension process and assessment of these operations that might inform understanding the entire process and outcome differences. Further, Duke’s chapter indicated that if only subskills are assessed, then struggling readers who have strong word identification skills but poor comprehension skills will not be identified.

Stahl, S. A., & Hiebert, E. H. (2005). The “word factors”: A problem for reading comprehension assessment. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates.

The authors of this chapter built an argument that fluent behavior (specifically reading rate, accuracy, and prosody) seems connected to the reader’s ability to comprehend a text. This relationship might be stronger for younger students or students with increased reading difficulty and impacts comprehension assessment: Specifically, if a student misses individual words, there might be implications. The student’s understanding of the coherent structure of the passage may be unaffected, but the questions that deal with literary response or those that require the reader to infer the answer through use of prior knowledge may be impacted. These are important implications for separate measures of decoding/fluency and comprehension.

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. Available at <http://www.readingassessment.info>

This report examines the characteristics of students within seven disability categories (specific learning disabilities, speech/language impairments, mental retardation, emotional/behavioral disabilities, autism, deaf/hard of hearing, and visual impairments). Skills deficits isolated to one component of reading are shown to be more likely in students from certain disability groups. Such a pattern is observed, for example, in students with autism spectrum disorder (comprehension), visual impairment or blindness (word recognition, fluency), and deafness or hard of hearing (phonemic awareness, word recognition). Evidence that a pattern of skill deficits isolated to one component of reading is observed in students with learning disabilities supports the assessment of each component in isolation so that a deficit in one component of reading

does not preclude showing proficiency in another component of reading.

Guideline 2-C. Use criteria to select texts that represent different genres and promote the use of interesting passages that are likely to engage all students.

Afflerbach, P. (2004). *National Reading Conference policy brief: High stakes testing and reading assessment*. Oak Creek, WI: National Reading Conference.

The author outlined a critique of current high stakes tests and noted, “The format of current high stakes reading tests limits our ability to know how student read critically, how they evaluate what they read and how they use the knowledge that they gain through reading” (p. 6). Afflerbach also commented on the texts that should be included in these reconsidered assessments: “Reading assessment should reflect...various texts and purposes” (p. 12).

Armbruster, B. B. (1994). The problem of inconsiderate text. In G. Duffy, L. Roehler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 202-217). New York: Longman.

This book chapter described how various disciplines, such as history and science, are composed of texts with structures that are different from each other. The author also discussed the role of these structures on students’ comprehension. Readers rely on distinct structural characteristics as they seek to understand what they read. In addition, a reader’s purpose for reading impacts how he or she reads texts with various text structures. Armbruster’s chapter noted that the way texts are organized shows that selections that are more coherent or considerate result in greater student understanding.

Babbitt Bray, G., & Barron, S. (2003-2004). Assessing reading comprehension: The effects of text-based interest, gender, and ability. *Educational Assessment*, 9(3 & 4), 107-128.

This study employed hierarchical linear models to investigate the relationship between students’ interest in reading texts and their performance on reading comprehension test items on those texts. Study participants included 19,735 students in grades 4-8 working with 98 different reading passages. Study findings produced a small but significant relationship between interest and performance on tests, which appeared to be stronger for girls and higher achieving students. A second hierarchical linear models analysis explored the relationship between certain passage characteristics and higher or lower student interest. The authors also discussed implications for test construction.

Boscolo, P., & Mason, L. (2003). Topic knowledge, text coherence and interest: How they interact in learning from instructional texts. *The Journal of Experimental Education*, 71, 126-148.

The authors’ primary interest was to extend previous research on the interactive effects of readers’ knowledge, interest, and text coherence. The participants were 303 students in the 10th and 11th grades. The students were grouped based on interest in topic (high

or low) and knowledge in topic (high or low). After reading a text, students engaged in text recall and post-test questions. Three texts were read: texts that had (a) minimal coherence, (b) local coherence, or (c) local and global coherence. Data were analyzed using multivariate analysis of covariance (MANCOVA). Results showed that students with high interest and high knowledge scored statistically significantly better on all texts. Results were slightly mixed for students with high interest/low knowledge, or low interest/low knowledge and overall, students with low interest/low knowledge scored the lowest. Findings indicated that when topic knowledge is low, interest can contribute to helping readers organize the text on a basic level but when topic knowledge is high and interest is high, a student can deeply process and understand text more easily.

deSousa, I., & Oakhill, J. (1996). Do levels of interest have an effect on children's comprehension monitoring performance? *British Journal of Educational Psychology*, 66, 471-482.

Several aspects of text have been shown to influence students' interest. Research also has indicated that text interest might be more important for low achieving students than for their counterparts. For example, deSousa and Oakhill found that for elementary students, text interest has a greater influence on reading comprehension for students who are poor comprehenders than for students who are considered good comprehenders. The authors stated that it is possible that lower achieving students are more likely to self-monitor their comprehension if they find the text interesting.

Dillon, D. R., O'Brien, D. G., Kato, K., Scharber, C., Kelly, C., Beaton, A., & Biggs, B. (2009). The design and validation of a motivating large-scale accessible reading comprehension assessment for students with disabilities. *Fifty-eighth yearbook of the national reading conference* (pp. 277-293). Milwaukee, WI: The National Reading Conference.

This article reported the findings of a calibration study where the goal was to develop a pool of highly engaging, motivating texts, many with color photos and illustrations, for a large scale comprehension assessment for 4th and 8th grade students with and without disabilities. A total sample of 1,245 students participated, including 627 students from intact classrooms in grades 3-5 and 618 students from intact classrooms in grades 7-9. After completing multiple choice items, students also rated how interesting and challenging the passages were. Students at both grade levels rated expository passages as more interesting than literary passages. Data were also examined to determine whether students' motivation (interest) affected their reading performance. Linear regression analysis was conducted to test the hypothesis that motivation is positively correlated with reading performance *especially* for low-performing students. Results indicated that for the lowest scoring 4th grade students (e.g., those scoring within the lowest quartile on the test), passage interest was positively correlated at the .05 level with performance on the test. This included passages that were expository and narrative in equal proportion (7 of the 10 expository and 7 of 10 literary were positively correlated). In addition, overall results for the lowest scoring 8th grade students (e.g., those scoring within the lowest

quartile on the test) indicated that passage interest was correlated at the .05 level with performance on about half of the passages. This included passages that were expository and narrative in equal proportion (6 of the 10 expository and 6 of 10 literary).

Guthrie, J. T., & Wigfield, A. (2005). Roles of motivation and engagement in reading comprehension assessment. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates.

This chapter dealt with the characteristics of an assessment that can influence student motivation and performance. Using findings from research studies, the authors argued that the characteristics of texts (e.g., the genre, topics, and text structures) and the opportunity for students to have a choice in selecting reading materials makes a difference in students' engagement and motivation to read and thus impacts their performance on assessments.

International Reading Association & National Council of Teachers of English. (1994). Standards for the assessment of reading and writing. Newark, DE: Authors. <http://www.reading.org/downloads/publications/books/bk674.pdf>

This document provided 11 standards to guide decisions about assessing the teaching and learning of reading and writing. The standards were fully described with rationales and implications, and case studies were provided to show best practices in the classroom. Standards 1 and 8 are important to the idea of selecting particular texts for students to read. Standard 1 stated, "The interests of the student are paramount in assessment" (pp. 10-11). Standard 8 noted, "The assessment process should involve multiple perspectives and sources of data" (pp. 20-21).

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1-43.

In this study, two experiments on text coherence were conducted using middle school students (11-15 years old). In the first experiment, which tested the authors' hypothesis that text coherence would differentially affect readers with high and low background knowledge, students were given a pretest and post-test and participated in a text recall. Students who read more coherent texts were able to recall more of the texts. In the second experiment, the authors sought to determine the relationship between students' background knowledge and the coherency of text. The authors predicted that the students with high background knowledge would benefit from a less coherent text, because of the added processing needed. Students were given a prior knowledge questionnaire and randomly assigned texts that were minimally or maximally coherent. The hypothesis held true—students with a strong background knowledge benefited more from the less coherent text, and conversely, students with low background knowledge benefited more from a more coherent text.

Mohr, K. (2006). Children's choices for recreational reading: A three-part investigation of selection preferences, rationales, and process. *Journal of Literacy Research*, 38, 81-104.

The purpose of this article was to determine first graders' preferences, rationales, and processes when choosing picture books to own. In the study, 190 first graders selected their favorite picture book from among nine high quality texts representing a variety of topics, media, and genres; 22 of these students were also interviewed. Conclusions from the study indicated that an overwhelming majority of the students selected informational books, especially animal books. Results contradict the authors' hypothesis that young students—particularly girls—would prefer narrative text. Students' selection rationales focused on topic or perceived content of the text.

National Assessment Governing Board. (2008). Reading framework for the 2009 National Assessment of Educational Progress. Washington, DC: Author. Available at: <http://www.nagb.org/publications/frameworks/reading09.pdf>

Pearson, P. D., Barr, R., Kamil, M. L., & Mosenthal, P. (Eds.), (1984). *Handbook of reading research* (Vol. I). Mahwah, NJ: Lawrence Erlbaum Associates.

Barr, R., Kamil, M. L., Mosenthal, P., & Pearson, P. D. (Eds.). (1996). *Handbook of reading research* (Vol. II). Mahwah, NJ: Lawrence Erlbaum Associates.

Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (Eds.). (2000). *Handbook of reading research* (Vol. III). Mahwah, NJ: Lawrence Erlbaum Associates.

Johnston's chapter in volume I of the *Handbook of Reading Research* indicated that unfamiliar texts should not be the focus of tests even to ameliorate the influence of prior knowledge. Johnston stated that when readers do not use their prior knowledge, this often results in poor comprehension. Over time, texts selected for reading tests were designed to describe task difficulty rather than reader ability. With the shift to selecting texts at particular reading levels of difficulty (independent; frustration), comprehensible text maintains readers' attention, so the interaction between text and reader is crucial. The research from Meyer and Rice, Purves, and Goldman and Rakestraw, described in volumes I–III of the *Handbook of Reading Research*, indicated how the structure of a text and the underlying ideas in the text are interrelated and relate a message to the reader that can impact the readers' ability to read, identify, and comprehend key ideas within the text. As Chall noted in the *Handbook of Reading Research* (Vol. II), readability is one way to access text difficulty, but readability deals with surface factors of prose. More recent reading research by Kintsch and Vipond has focused on text structure variables in theory construction, specifically the structure of the text represents the relationships among the ideas in the text.

Pearson, P. D., & Camperell, K. (1994). Comprehension of text structures. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of*

reading (4th ed., pp. 448–468). Newark, DE: International Reading Association.

This handbook chapter described research that indicated that the characteristics of informational and literacy texts are strikingly different and that readers focus on different aspects of the text as they read it. In addition, there is evidence that text structure does impact a reader's ability to comprehend a text and that students need exposure to, instruction in, and assessment of these various structures.

Sadoski, M. (2001). Resolving the effects of concreteness on interest, comprehension, and learning important ideas from the text. *Educational Psychology Review*, 13, 263-281.

The author reviewed recent studies on the effects of concreteness and mental imagery on text comprehension, interest, and memory. The research found that concreteness (as defined by language rich in connections) is a powerful predictor of comprehensibility and immediate recall, particularly when the topic is familiar to the reader. The research also indicated that concrete material tends to be more memorable and more interesting and engaging than more abstract material. Sadoski also examined research on seductive details, defined as, “novel, active, concrete and personally engaging but irrelevant information” (p. 272). He concluded that seductive details do not detract from recall of important ideas in a well-structured and coherent text.

Salinger, T., Kamil, M. L., Kapinus, B., & Afflerbach, P. (2005). Development of a new framework for the NAEP reading assessment. In B. Maloch, J. V. Hoffman, D. L. Schallert, C. M. Fairbanks, & J. Worthy (Eds.), *54th yearbook of the national reading conference* (pp. 334-348). Oak Creek, WI: National Reading Conference.

The NAEP 2009 framework outlined a rationale for the texts that will be used in the new assessment. The framework document stated, “Reading passages are selected to be interesting to students nationwide, to represent high-quality literary and informational material, and to be free from bias” (p. 1).

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading*, 3, 257-279.

This paper compiled several reviews of the literature on the role of text and interest. One review that examined 22 studies found an average correlation of .27 ($p > .01$) between personal interest and text learning. The review also indicated that the positive relation between interest and text learning was independent of the following factors: text type (genre), difficulty, length of the passage, grade level or the reading ability of the students. Most of the studies in this review did not account for the amount of time students used to read passages. In an additional analysis of 14 studies, students judged the “interestingness” of the texts they were asked to read. These ratings were then correlated with the students' text comprehension scores. Findings indicated that deep processing of text was higher for those texts that students rated as more interesting. Increased deep processing was related to increased comprehension scores. This analysis of multiple

studies suggests that “it is reasonable to conclude that there are situations or domains in which learning depends on motivational factors, such as interest, independently of cognitive ability or prior knowledge” (p. 272).

Shallert, D. L., & Reed, J. H. (1997). The pull of text and the process of involvement in reading. In J. Guthrie & A. Wigfield (Eds.), *Reading engagement: Motivating readers through integrated instruction* (pp. 68-85). Newark, NJ: International Reading Association.

This chapter described what makes a text engaging to a student. For example, authors Schallert and Reed found that students were more involved when a text included vignettes, illustrated examples of ideas, had characters that students could identify with, said unexpected things or had unexpected events, was personally relevant to students, or focused on “life themes...[such as] death, danger,...power, and destruction” (p. 73). The authors noted that to engage students, a text can be neither too easy nor too challenging, but needs to pull the reader in with certain textual moves that readers find appealing.

Wade, S. E., Buxton, W. M., & Kelly, M. (1999). Using think-alouds to examine reader-text interest. *Reading Research Quarterly*, 54, 194-216.

In this study on the role of interest in reading, college students read a text and then completed a think-aloud activity, which was used to analyze data. Participants cited the following characteristics as most important in increasing their interest in a text: importance or value, reader’s connections, and author’s connections. The most frequently cited characteristic that decreased interest was lack of comprehensibility. The authors of this article also provided a thorough review of literature on the effect of interest on text recall, including the work of researchers such as Alexander, Kulikowich, and Jetton; Alexander and colleagues; Beck and colleagues; Garner and colleagues; Goetz and colleagues; Harp and Mayer; Sadoski and colleagues; Sadoski and Quast; Schraw and colleagues; Wade and Adams; and Wade and colleagues.

Guideline 2-D. Use criteria to determine the visual elements that should be included within texts while avoiding distracting elements that impact students’ comprehension.

Allman, C. B. (2004). *Test access. Making tests accessible for students with visual impairments: A guide for test publishers, test developers, and state assessment personnel* (2nd ed.). Louisville, KY: American Printing House for the Blind.

This guide, which was written specifically for test developers, emphasized the importance of using criteria to determine the visual elements that should be included within texts and provided recommendations for ensuring that visual elements are accessible to students who are blind or visually impaired. The guide stated, “It is recommended that as much information as possible be included in the text of a test item. This will help prevent the introduction of pictures that contain information necessary for the selection of the

correct answer, but which cannot be adequately brailled, presented in large print, or described in audio format” (p. 8).

Barr, R., Kamil, M. L., Mosenthal, P. & Pearson, P. D. (Eds.). (1996). *Handbook of reading research* (Vol. II). Mahwah, NJ: Lawrence Erlbaum Associates.

Chall and Squire’s chapter on the publishing industry and textbooks indicated that studies have been spotty in the area of the role of visual design and reading comprehension. The authors maintained that research over the previous 70 years has indicated that the page layout and type size can make a difference in reading efficiency and comprehension. This research has also indicated that illustrations can either facilitate or hinder comprehension, depending on the visuals used (components of the visual) and the level of the reading material. Waller’s chapter on typology and discourse reviewed the literature on the role of the design and arrangement of printed text, including graphic design, on reading comprehension. The review included the work of scholars such as Miles Tinker, who did considerable research on variables such as type size, type design, and the color of ink and paper. But critics of the research on the effect of typology on reading have noted that the experimental results often showed small differences, thus little legibility research has transpired in reading in recent years. Other researchers have found that reader preferences in typographical style impact reading speed (indicated in the work of Burt, Cooper, & Martin). Waller’s study also reviewed research on graphic designs (which is called topic structure) because it displays structures and boundaries within the discourse, such as graphic organizers, and visual informativeness, which is called access structure, because it provides visual clues to aid the reader. Waller’s chapter also noted that the research is mixed as to whether and how visual elements facilitate readers (e.g., immediate recall is enhanced; delayed may not be).

Brookshier, J., Scharff, L. F. V, & Moses, L. E. (2002). The influence of illustrations on children’s book preferences and comprehension. *Reading Psychologist*, 23, 323-339.

The purpose of this study was to determine whether illustrations, in addition to text, lead to better comprehension, as well as to investigate the preference of first and third graders for types of illustration. Students were assigned to text-only, pictures-only, or pictures-and-text conditions. The pictures also varied by being realistic or abstract. Both first and third graders showed the highest level of comprehension for the pictures-and-text condition and the lowest level of comprehension for pictures-only condition. All students showed a preference for realistic pictures.

Levie, W. H., & Lentz, R. (1982). Effect of text illustrations: A review of research. *Educational Communication and Technology*, 30, 195-232.

Willows, D. M., Borwick, D., & Hayren, M. (1981). The content of school readers. In G. E. Mackinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 2, pp. 100-179). New York: Academic Press.

The Levie and Lentz and the Willows, Borwick, and Hayren reviews of research literature both indicated comprehension was superior with illustrated texts. Results from reviewed studies showed that students' comprehension was greater when they read texts that included illustrations as opposed to when these texts appeared without illustrations. In particular, younger students performed better in their reading comprehension using illustrated texts.

Rose, T. L. (1986). Effects of illustrations on reading comprehension of learning disabled students. *Journal of Learning Disabilities, 19*, 542-544.

This article highlighted the importance of using criteria to determine whether illustrations should be retained or excluded, based on whether they help or hinder students' comprehension of text. The purpose of this study was to investigate the effects of illustrations on the reading comprehension of elementary school students with learning disabilities. The 32 subjects in the study, who were between 9–12 years old, each read two randomly assigned stories (one with illustrations and one without illustrations). Following each story, they answered comprehension questions. Results indicated that students with learning disabilities comprehended stories without illustrations significantly better than stories with illustrations. No students demonstrated better comprehension with the stories that were illustrated, although two students performed equally well in both conditions.

Sanchez, C. A., & Wiley, J. (2006) An examination of the seductive details effect in terms of working memory capacity. *Memory and Cognition, 34*, 344-355.

These studies indicated that test developers need to carefully examine illustrations that accompany texts, retain those that enhance comprehension, and consider removing those that draw students' attention in directions that are unproductive. The purpose of this experiment was to see if differences in working memory capacity would predict the seductive-detail effect and impact the processing of illustrated text. A total of 72 undergraduates participated in the study. After reading the passage, students were tested on their comprehension of the text. The texts used in the study had components with no illustrations, seductive illustrations, or conceptual illustrations. Results indicated that students with low working memory capacity were "seduced" by the irrelevant illustrations and for students with high working memory, seductive illustrations seemed to have the opposite effect (irrelevant information appeared to help them learn). In the second experiment, students' reading patterns were examined while they read the seductively illustrated text. This second part of the experiment confirmed the results of the first experiment.

Guideline 2-E. Present reading tasks that students perceive as worthwhile and that enable them to be self-efficacious as learners.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

This foundational book outlined key ideas relating to self-efficacy: “Such beliefs influence the courses of action people choose to pursue, how much effort they put forth in given endeavors, how long they will persevere in the face of obstacles and failures, their resilience to adversity, whether their thought patterns are self-hindering or self-aiding, how much stress and depression they experience in coping with taxing environmental demands, and the level of accomplishments they realize” (p.3). These ideas are key to creating accessible reading assessments. For example, if tasks are created that students perceive to be doable, interesting, and worthwhile, and if they have some choice in test activities, then they may be more likely to persevere and complete the assessment—even when faced with challenging passages and items.

Campbell, J. R., & Donahue, P. L. (1997). *Students selecting stories: The effects of choice in reading assessment*. Washington, DC: National Center for Education Statistics.

The researchers examined the feasibility and measurement impact of offering test takers in the 8th and 12th grades a choice of reading material on an assessment of reading comprehension. While other studies indicate that the choice option can support students’ self efficacy, this study suggested that caution needs to be taken when using the choice condition. In the design, a group of readers who could select from among seven stories to read as part of the 1994 NAEP assessment was compared with a group who were assigned stories. Researchers found no significant effect for choice for 12th graders and slightly lower performance in the choice condition for 8th graders. Results may have been influenced by the fact that subjects only had a certain amount of time to both select passages to read and read these passages and answer questions—possibly impacting comprehension results. Subjects also only read narrative texts versus a mixture of expository and narrative materials on the assessment.

Educational Testing Service. (2005). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Author.

In the guidelines, test developers are advised to allow test takers to choose an appropriate task, a response mode, or the conditions of testing whenever possible. The skill being measured will determine whether an option of choice is permissible. According to the guidelines, “The test takers are likely to perceive the assessment as fairer if they have these choices” (p. 5).

Guthrie, J. T., Van Meter, P., Hancock, G., Alao, S., Anderson, E., & McCann, A. (1998). Does concept-oriented reading instruction increase strategy use and conceptual learning from text. *Journal of Educational Psychology*, 90, 261-278.

This study reported that the reading comprehension of elementary school students increased when they read informational texts—many self-selected—within the concept units focused on science topics. Students worked on meaningful tasks with other peers and independently. Results found that students viewed the tasks as meaningful and worthwhile, used reading strategies effectively to obtain information, felt self-efficacious

in their reading abilities, and were highly engaged in their work.

Miranda, A., Villaescusa, M. I., & Vidal-Abarca, E. (1997). Is attribution retraining necessary? Use of self-regulation procedures for enhancing the reading comprehension strategies of children with learning disabilities. *Journal of Learning Disabilities, 30*(5), 503-512.

The authors sought to learn more about the effect of self-regulation instruction procedures on the cognitive and metacognitive reading development of students with learning disabilities, as well as the effect of explicit attribution retraining. The research was conducted in five different schools in the Alicante region of Spain, using experimental and control groups of students with and without learning disabilities. All students were pre- and post-tested, and the experimental groups received instruction in self-regulated learning strategies and attribution retraining. The authors found that self-regulated procedures were successful in increasing students' reading comprehension strategies: The experimental group of students with learning disabilities performed on the post-test at the level of students without learning disabilities. Students with learning disabilities in the control group maintained their differences. Results suggest that helping a reader develop self-regulation is key to enhancing the belief that he or she is self-efficacious and can complete tasks successfully. Thus, readers with disabilities may be more willing to engage in challenging reading tasks and successfully complete them.

Paris, S. G., & Carpenter, R. D. (2004). Children's motivation to read. In J. V. Hoffman & D. L. Schallert (Eds.), *The texts in elementary classrooms* (pp. 61-82). Mahwah, NJ: Lawrence Erlbaum Associates.

The authors of this report stated that there is no single way to describe motivation. They argued for a situational model where "motivation does not reside only in the child, the text, the task, the home or the classroom...but that students are more or less motivated for reading depending on the interaction of all of these factors" (p. 79). The authors also stated that teachers need to be concerned with students' feelings about reading while also seeking to raise their skill levels: "Motivation and self-regulated reading must be as important as test scores" (p. 79). They claimed that assessments can also be created that are organized by the principles of "constructing meaning with choice, control, collaboration, challenge, and consequences that enhance self-efficacy" (p. 75).

Shell, D. F., Murphy, C. C., & Bruning, R. H. (1989). Self-efficacy and outcome expectancy mechanisms in reading and writing achievement. *Journal of Educational Psychology, 81*(1), 91-100.

In their study of undergraduate students in educational psychology classes, the authors investigated the relationship between self-efficacy and outcome expectancy beliefs and achievement in reading and writing. Students took the Degrees of Reading Power test and provided writing samples, which were analyzed holistically. Regression analysis of the study results indicated that beliefs were significantly related to performance in

reading and writing. Substantial variance in reading achievement was accounted for by self-efficacy beliefs.

Turner, J. C. (1995). The influence of classroom contexts on young children's motivation for literacy. *Reading Research Quarterly, 30*, 410-441.

The purpose of this study was to understand the importance of classroom activities on students' motivation to read and their reading achievement. The researcher examined all teachers' reading activities and tasks and coded them as either open or closed tasks. Open tasks allowed students more control over choices involved within an activity (e.g., the choice of a text to read) and how they might complete the reading task (e.g., writing a response or creating a project to show comprehension of the text). Closed tasks were completely designed by the teacher and offered no student choice or ability to determine how a task might be completed. Students who engaged in open-ended tasks were more self-efficacious and willing to take on challenging reading passages and tasks because they had choice in the task design and outcome.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159-195.

This article summarized results from tests that have allowed examinee choice. The article provides an extensive discussion of choice including a history of the use of choice for academic assessments as well as a discussion of the complexity of the psychometric issues associated with choice. The authors make the point that it is difficult to ensure comparability of scores obtained on assessments when examinees exercise choice of tasks. In addition, they refer to results of studies that have shown examinees and subgroups of examinees are sometimes disadvantaged by choices they make on examinations. The authors conclude that building choice into a test is possible, but, "... it requires extra work. Approaches that ignore the empirical possibility that different items do not have the same difficulty will not satisfy the canons of good testing practice, nor will they yield fair tests. But, to assess the difficulty of choice items, one must have responses from an unselected sample of fully motivated examinees. This requires a special sort of data gathering effort" (p. 190).

Wigfield, A., & Eccles, J. S. (2001). (Eds.). (2002). *Development of achievement motivation*. New York: Academic Press.

This foundational book presented the theory and research on how motivation changes as children progress through school, gender differences in motivation, and motivational differences as an aspect of ethnicity. The book included edited chapters that report research studies and findings including central constructs and theories such as self-efficacy, expectancies and values, self-worth, beliefs about the nature of ability, achievement goals, intrinsic motivation, and self-regulation. Results indicated that students who feel that they have the skills and strategies to succeed (competence related beliefs) become more task-engaged and this impacts their achievement in favorable ways.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, 29(3), 663-676.

Zimmerman, Bandura, and Martinez-Pons examined parents' academic goals linked to children's personal goals, which in turn were linked to their academic goals. The authors in this study of 9th and 10th grade social studies students found a strong causal path between self-efficacy and academic achievement, using a questionnaire and final grades as their data sources. Not only did students with higher self-efficacy achieve higher year-end grades ($r = .41$), but they also set higher grade goals ($r = .51$). The authors stated that while certain aspects of student achievement remain unexplained, "self-efficacy and goals in combination contribute to subsequent academic attainment" (p. 674).

Guideline 2-F. Ensure that test blueprints are aligned with the state standards.

American Educational Research Association. (2003). Standards and tests: Keeping them aligned. *Research Points*, 1(1), 1-4.

This issue of *Research Points* provided guidelines for school districts to use to determine if their assessments are aligned with the state standards. Areas of suggested consideration included content match, breadth of coverage and balance across standards, level of challenge in students' response mode, and weeding out of extraneous test material not covered in the standards.

Baker, E. L. (2004). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform* (CSE Tech. Rep. No. 645). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Baker used a series of metaphors to discuss the importance of alignment in results-based reform (RBR). She explored alignment of an entire educational system including outcomes, processes, and goals. She emphasized the importance of a coherent curriculum, "...one that emphasizes both broadly and concretely the intentions of the standards, and the content and skills to be taught and learned" (p. 15). In concluding the article, she made the point that alignment should be treated as a goal rather than a bureaucratic requirement to be checked off. Baker wrote, "If it is the latter, we will never attain it. And should we believe we have 'aligned' our system, we must remember that the world moves, and alignment strengthens and weakens with change" (p. 20).

Johnstone, C. J., Moen, R. E., Thurlow, M. L., Matchett, D., Hausmann, K. E., & Scullin, S. (2007). *What do state reading test specifications specify?* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

This paper reviewed state assessment blueprints and specifications from 49 states. Results indicated that states' reading items were primarily focused on literal comprehension of reading passages and analyzing text (inferential comprehension). To a lesser extent, states included foundational skill items on assessments. The majority of state assessment

items were multiple-choice items on tests. Results from this study were compared with another study by Johnstone, indicating that some standards (e.g., critical analysis of text or connecting literature to one's own life) are very difficult to test on state assessments.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7. Retrieved February 20, 2009, from <http://PAREonline.net/getvn.asp?v=7&n=21>

This article argued that the alignment of standards and assessments is an essential element of high quality assessment practice in the era of No Child Left Behind legislation. The author outlined processes for aligning assessment and state standards, citing Webb (2006) as his main source (see below).

National Assessment Governing Board. (2008). *Reading framework for the 2009 National Assessment of Educational Progress*. Washington, DC: Author. Available at: <http://www.nagb.org/publications/frameworks/reading09.pdf>

This document provided readers with an example of a well-executed reading framework. Specifically, the document provided an overview of the specifications used for developing 2009 NAEP reading assessments. One particularly noteworthy distinction between NAEP items and statewide items is that NAEP assesses the foundational skills (e.g., word recognition, phonics) as these are embedded within comprehension and analysis/interpretation items. This approach differs from some states that opt to include items in tests that specifically test foundational skills.

Niemi, D., Baker, E. L., & Sylvester, R. M. (2007). Scaling up, scaling down: Seven years of performance assessment development in the nation's second largest school district. *Educational Assessment*, 12(3&4), 195-214.

This article showcased the importance of connecting performance assessment at every level to explicit learning goals and state standards. The authors summarized a 7-year performance assessment collaboration between assessment researchers and the nation's second largest school district. The project was designed to test assessment design models and scoring procedures involving more than 300,000 students per year. Findings were reported to reflect foundations for large-scale assessment, capacity building, score reliability, and professional development.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003/2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, 9(1&2), 1-27.

This article examined the relationship between state standards and assessments that measure what the standards prescribe. The authors illustrated a methodology that checks alignment of tests to state standards by examining patterns of the degree of alignment in a small but representative sample of states. The study found that, while individual items align quite well with some standards, the examined tests as a whole are not well-aligned. According to the authors, such misalignment can have serious consequences for instruction and validity of test results.

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155-180). Mahwah, NJ: Lawrence Erlbaum Associates.

This chapter provided guidelines for test developers on how to develop test specifications that align with state standards for content knowledge and skills. The author emphasized the importance of clear content specifications and discussed four criteria for framing content specifications. These include categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-representation. In addition this chapter outlined processes for selecting items for a particular test. One process that is common for this task is developing a two-way matrix whereby the rows represent content areas and columns represent skills to be measured by the test. From there, test developers may generate representative items for each combination of skill and content.

Principle 3: Reading assessments are developed with accessibility as a goal throughout rigorous and well-documented test design, development, and implementation procedures.

Guideline 3-A. Initial test design considers the characteristics of all test takers.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

One chapter in this document of standards for testing is titled “Fairness in Testing and Test Use.” It stated, “There is general consensus that consideration of bias is critical to sound testing practice” (p. 74). Several of the standards in that chapter and in other chapters asserted the importance of identifying and eliminating inappropriate item bias as tests are developed and used. Bias was defined as a technical term and “it is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of identifiable subgroups” (p. 74). This points to the necessity of developing assessment practices that permit accurate inferences to be drawn for all test takers.

Hansen, E. G., & Mislevy, R. J. (2008). *Design patterns for improving accessibility for test takers with disabilities* (ETS Research Rep. No. RR-08-49). Princeton, New Jersey: ETS.

This report examined the use of design patterns for clarifying issues related to accessibility features for individuals with disabilities—such as low-vision and blindness—who are taking a test of reading. Design patterns are task-design aides (templates) that were originally developed with National Science Foundation support for assessing science inquiry skills. This report adapted design patterns for clarifying how variable features of a design for a reading comprehension test need to be matched to disability-

related characteristics of test takers in order to ensure accessibility. For example, the font size (a variable feature of the task situation) needs to be matched to the disability-related characteristic (e.g., low vision) of the test taker. An implication is that the use of design patterns may improve the validity and fairness of tests, as well as their accessibility for individuals with disabilities.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

These authors presented a framework for designing assessments, arguing that all assessments should rest on the three pillars of cognition, observation, and interpretation. They stated that research-based descriptions of how children develop understanding of particular subject matter should be foundational in test design. This contrasts with more purely statistical or psychometric approaches that start with large pools of items and winnow down to only those items that show desirable statistical characteristics. Instead, by having a clear model of what is to be measured and an intentional design for each item that is developed, test developers can plan from the outset the best ways of revealing the knowledge and skills of students with different characteristics.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger.

According to Schmeiser and Welch, "...the most important stage in the development of an educational achievement test is the design stage." One of the key decisions to be made in this first stage is "defining the intended examinee populations." The authors cited the importance of identifying and removing item bias through human judgment during fairness review procedures and by statistical analysis during data analysis procedures.

Thurlow, M. L., Quenemoen, R. F., Lazarus, S. S., Moen, R. E., Johnstone, C. J., Liu, K. K., Christensen, L. L., Albus, D. A., & Altman, J. (2008). *A principled approach to accountability assessments for students with disabilities* (Synthesis Rep. No. 70). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

In considering the qualities that make up an inclusive system of assessment used for accountability, one of the characteristics the authors identify is, "All assessments are designed from the beginning with a focus on accessibility for all students" (p. 7). This requires "knowledge of needs of the full range of students to be tested along with careful scrutiny of intended constructs and design of assessments" (p. 7). One help in ensuring that these needs are understood is the recognition that "stakeholders representing varied student subgroups are essential partners in shaping the development of assessment systems that appropriately address varied learner characteristics" (p. 5).

Guideline 3-B. Item development and evaluation considers the characteristics of all test takers.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Of the many standards suggested for educational and psychological tests in this document, Standard 3.6 attended to the development and evaluation of items and tasks and specifically addressed different groups of test takers and qualifications of item reviewers. It stated: “The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented” (p. 44). Standard 10.4 indicated that “modifications as well as the rationale for the modifications should be described in detail in the test manual and evidence of validity should be presented whenever available” (p. 106). This implies that diverse groups should be considered in the development of items and that rationales for modifications should be documented.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Tech. Rep. No. 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

This study used a quasi-experimental design to test the effects that a combination of culturally sensitive items, larger print, concise and readable language, and unlimited time had on student assessment scores. Effects were found to be positive, with students with disabilities and English language learners benefiting in particular. A content expert judged that there was a one-to-one correspondence of item constructs between the universally designed tests and the traditional tests. The study found statistically significant differences between mean scores on universally designed and traditional tests (effect size = .39). Student interviews indicated that timing, large fonts, and comprehensible language were most important to student comprehension of items. An implication is that expert item review and other methods are critical in establishing the validity of results of tests that have been modified with accessibility features. The author further noted, “Currently, most constructs are undefined by item designers” (p. 24), underscoring the need for precise definitions of constructs.

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

This article reviewed three approaches that can add validity evidence to states' item review processes. The first process was a structured sensitivity review process that focused on universal design considerations for items. The second method was a series of statistical analyses intended to increase the limited amount of information that can be derived from analyses on low-incidence populations (such as students who are blind or deaf or have cognitive disabilities). The third approach was think-aloud methods, which were described as a way to understand why particular items might be problematic for students. An implication is, as stated by the authors, that "a potential next step to this article is to use all item review methods (expert review, quantitative analysis, and think aloud techniques) collectively in a state's item development processes" (p. 35).

Koenig, J. A., & Bachman, L. F. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*. Washington, DC: National Academies Press.

This report from the Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large Scale Assessments addressed critical issues in the assessment of English language learners and students with disabilities, especially for the National Assessment of Educational Progress (NAEP), but relevant to other assessments as well. In Recommendation 6.1, Koenig and Bachman stated: "NAEP officials should identify the inferences that they intend should be made from its assessment results and clearly articulate the validation arguments in support of those inferences" (p. 122). Among the diverse methods for gathering evidence for these arguments are those that examine "test takers' cognitive processes" (Recommendation 6.2, p. 122), of which think-aloud protocols are an instance. Chapter 6 of this report provided an example of how to explicate the argument for the validity of assessments given under altered conditions. An implication of this report is that one should document the intended interpretations of tests and the validation arguments to support those interpretations.

Guideline 3-C. Test assembly and evaluation considers the characteristics of all test takers.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

This document from three national organizations included Standard 10.5, which addressed the need to ensure that inferences from tests reflect the intended construct rather than students' disabilities. This standard stated: "In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement" (p. 106). Standard 10.3 called for direct involvement of individuals with disabilities in the piloting of tests: "Where feasible, tests that have been modified for use

with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications” (p. 106). The implication is that the development and formative evaluation of tests should be focused on maintaining validity and should involve test takers with disabilities.

Moen, R. E., Liu, K. K., Thurlow, M. L., Lekwa, A., Scullin, S., & Hausmann, K. (2009). Identifying less accurately measured students. *Journal of Applied Testing Technology*, 10(2).

The purpose of this study was to examine the prospect of using teachers to identify students whose scores on typical reading tests would be misleadingly low. A small number of teachers were asked to identify such students and to explain why each student had been identified. Researchers interviewed the teachers, examined evidence the teachers offered in support of their judgments, and interviewed and assessed students. Based on this in-depth assessment, the researchers accepted teachers’ assertions for 14 of 20 identified students. The concluding discussion suggested procedures that could involve teachers more effectively in identifying assessment factors that reduce the accuracy of reading tests for some students. This study provided an example of how to evaluate an assessment to ensure that students with diverse characteristics are accurately assessed.

New England Compact. (2007). *Reaching students in the gaps: A study of assessment gaps, students, and alternatives* (Grant CFDA #84.368 of the U.S. Department of Education, Office of Elementary and Secondary Education, awarded to the Rhode Island Department of Education). Newton, MA: Education Development Center.

In this study, researchers compared scores obtained on a state math test with teachers’ judgments about students’ math achievement. Particular attention was paid to Gap 1 students, who performed poorly on the state math test despite performing proficiently in the classroom. Researchers asked teachers for their opinions about what aspects of assessments prevent Gap 1 students from showing on tests what they can do in the classroom and proposed assessment adjustments that might help such students show what they know on tests. This study provided an example of how to evaluate an assessment to ensure that students with atypical performance characteristics are accurately assessed.

Valencia, S. W., & Buly, M. R. (2004). Behind test scores: What struggling readers *really* need. *Reading Teacher*, 57(6), 520-531.

This article looked beyond reading test scores to distill specific reading abilities of students who failed a standardized reading test. The research was conducted in a northwestern U.S. school district of 18,000 students, approximately half of which had failed the state fourth-grade reading test. Of the students who had scored below standard, 108 of them participated in individual reading assessments administered in a one-on-one format for approximately 2 hours over several days. The findings were

stratified into three categories of reading abilities: word identification, meaning, and fluency. The researchers observed individual differences in these categories, which hampered student performance on state assessments.

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

This frequently cited paper described a process used by item reviewers in mathematics and science to determine the level of alignment between standards and assessments. Webb's model included rating items on their categorical concurrence (between standards and assessments), depth of knowledge consistency, range of knowledge correspondence, and balance of representation. Although this paper examined math and science, the process may be used as a blueprint for alignment studies of reading items. An implication of this work is that alignment studies can help evaluate the content validity of assessments, whether or not they have been modified for individuals with disabilities. Such studies can also help guide the assembly of items into tests and subtests that can result in scores for either an overall reading construct or a reading component skill.

Guideline 3-D. Document the steps that have been taken to ensure that the characteristics of all test takers have been considered.

Allan, J. M., Bulla, N., & Goodman, S. A., (2003). *Test access: Guidelines for computer administered testing*. Louisville, KY: American Printing House for the Blind.
Available at: <http://www.aph.org/tests/access/access.pdf>

This publication addressed issues that may arise for test takers who have disabilities, particularly those with low vision or who are blind. In spite of advances in technology (e.g., computer-administered tests), there is still an achievement gap between students with and without visual impairments. The authors presented the principles of inclusive design and argued that tests must be made accessible to all potential test takers, regardless of format or disability. They further recommended, "Developers of computer-based tests should also document those methods and procedures that are employed when striving toward accessibility." The recommendation regarding documentation is consistent with a goal of explicating, preserving, and sharing of evidence supporting intended interpretations and uses of assessments with accessibility features.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

To ensure the appropriateness of accommodation assignments, Standard 10.10 of these joint standards for tests stated, "A test professional needs to consider reasonably available information about each test taker's experiences, characteristics, and capabilities that might impact test performance, and document the grounds for the modification"

(pp. 107-108). By thus documenting the grounds for any modifications that have been made in the assessment, the test professional provided information that anyone who interpreted results from such modified assessments may need to make accurate inferences.

Principle 4: Reading assessments reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student's proficiencies.

Guideline 4-A. Begin the assessment development or revision process by reviewing allowed accommodations to determine whether they could be incorporated into the design of the assessment.

Koenig, J. A. (Ed.), (2002). *Reporting test results for students with disabilities and English-language learners*. Washington, DC: National Research Council.

This report was a summary of a workshop addressing the reporting of test results. It covered a wide range of issues, including a call for better test design, in the chapter that synthesized issues and direction for future study. Specifically mentioned by workshop discussants and presenters was the consideration of ways to “construct tests from the outset to minimize the effects of and the need for accommodations” (p. 71).

Thompson, S. J., Thurlow, M. L., & Malouf, D. (2004, May). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology, 1*. Available from [http:// www.testpublishers.org/journal.htm](http://www.testpublishers.org/journal.htm)

This article presented a set of elements considered by the authors to be essential in creating universally designed assessments, which are assessments designed to be appropriate for the widest range of students without changing the construct that the assessment is intended to measure. Of the seven elements that the authors proposed, Element 4 addressed the need for assessments to be amenable to accommodations. Examples were provided of ways to think about incorporating accommodations into the design of the assessment, including consideration of features of the assessment that may need to be adjusted to facilitate incorporation (e.g., use of vertical or diagonal text may not be appropriate if students are using screen readers or braille versions of assessments).

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653-673). Mahwah, NJ: Lawrence Erlbaum Associates.

The chapter included a detailed discussion of how accommodations may be used to provide special needs populations with access to tests as well as a discussion of the major challenges that surround the use of accommodations and modifications. The chapter addressed universally designed assessments, suggesting that there is a need for well-designed tests that minimize the need for accommodations. It also argued that “well-

designed assessment is a better measure for all students, including those students with disabilities and those who are learning English” (p. 668).

Guideline 4-B. Identify and determine the essential accommodations that are still needed after incorporating as many as possible into the assessment.

Loeding, B. L., & Greenan, J. P. (1998). Reliability and validity of generalizable skills instruments for students who are deaf, blind, or visually impaired. *American Annals of the Deaf*, 143(5), 392-403.

This study provided a discussion of how different types of testing accommodations impact test reliability over several content domains and testing accommodations. The study included 103 deaf and blind students enrolled in secondary vocational classes who each took at least one of four assessments (interpersonal relations skills, mathematics skills, reasoning skills, or communication skills). For each of the four domains, a self-rating, a teacher rating, and a performance assessment specifically designed for vocational students was given. Item formats varied and included multiple-choice, open-ended, matching, and interactive tasks such as following directions or giving directions. Although test-retest, reliability, and validity ranged from moderate to high, the small numbers of students (30 or less for each measure) was low. The discussion noted accommodations that could be built into the assessment (such as videotaped signing of item content and combination audio CD and large print).

Thompson, S. J., Quenemoen, R. F., & Thurlow, M. L. (2006). Factors to consider in the design of inclusive online assessments. In M. Hricko & S. L. Howell (Eds.), *Online assessment and measurement: Foundations and challenges* (pp. 102-117). Hershey, PA: Information Science Publishing.

This chapter presented factors to consider in the design of online assessments so that they are appropriate for all students, including students with disabilities and English language learners. Addressing assistive technology, the authors make the point that regardless of the format of the assessment and how well universally designed it is, there is a need to continue to provide accommodations, including assistive technology. They also pointed out that adaptations may be needed in the assistive technology to ensure that all students are able to use it, even if it is well-incorporated into the design of the assessment.

Guideline 4-C. Develop a strong rationale and evidence to support the validity of inferences from assessment results when accommodations are provided.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

The standards in this document address use and evaluation of testing practices. Standard 10.4 stated, “If modifications are made or recommended by test developers for test takers

with specific disabilities, the modifications as well as the rationale for the modification should be described in detail in the test manual and evidence of validity should be provided whenever available. Unless evidence of validity for a given inference has been established for individuals with the specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores” (p. 106).

Hansen, E. G., Mislavy, R. J., & Steinberg, L. S. (2008). *Evidence-centered assessment design for reasoning about testing accommodations in NAEP reading and mathematics* (ETS Research Rep. No. RR-08-28). Princeton, NJ: ETS.

This study described extensions to evidence-centered assessment design (ECD) for reasoning about the impact of accommodations and other accessibility features (e.g., universal design features) on the validity of assessment results. The paper gave several examples from NAEP reading and mathematics. This study also explored the use of Bayes nets (also called *belief networks*) for carrying out such analyses. The study found that ECD-based techniques may be useful in analyzing the effects of accommodations and other accessibility features on validity. An implication is that such design capabilities may increase assessment designers’ capacity to employ accessibility features without undermining validity.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Among the many guidelines in this document to encourage fair testing practices, Guideline C2 in the *Code of Fair Testing Practices in Education* stated that test developers should “provide guidance regarding the interpretations of results for tests administered with modifications” and “inform test users of potential problems in interpreting test results when tests or test administration procedures are modified” (p. 8).

Kane, M. T. (2006). Validation. In R. L. Brennan, *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

In this chapter about the validation of interpretations and uses of educational measurements, the author noted, “An important class of exceptions to the applicability of standardized testing procedures and therefore standard interpretive arguments involves examinees with disabilities” (p. 28). He also stated, “The goal is to reach the same kind of conclusions for all students, and the testing accommodations are designed to achieve this goal” (p. 28). The implication of this chapter is that accommodations and other accessibility features entail an alteration to the standard interpretive argument, and the argument should be explicated and documented.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, 71(1), 53-104.

This article presented a summary of the issues surrounding testing accommodations for students with disabilities. The article opened with highlighting the history of legislation related to assessment of students with disabilities and summarizing guidelines determined by testing organizations. The second section of the article described disabilities accommodated in large scale assessments and emphasized the role of learning disabilities in student assessment. The article then summarized legal cases associated with testing accommodations and described psychometric issues as well as results and ideas for future research. Finally, the authors examined the emerging role of learning disabilities in the area of testing accommodations and addressed the issues of equity and fairness in assessment.

Thurlow, M. L., Quenemoen, R. F., Lazarus, S. S., Moen, R. E., Johnstone, C. J., Liu, K. K., Christensen, L. L., Albus, D. A., & Altman, J. (2008). *A principled approach to accountability assessments for students with disabilities* (Synthesis Rep. No. 70). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

The authors of this report indicate that an important characteristic of a principled approach to inclusive systems of assessment for accountability is, “Accommodation policies are informed by the construct to be measured, available research findings, and the purpose of the assessment” (p. 8). They say that test users need to set “policies that indicate which changes in testing materials or procedures can be used during assessments, under which conditions, and whether the use of the accommodations or modifications might have implications for scoring or aggregation of scores” and that these policies “may change by student characteristic” (p. 8). An important part of this process is “to gather stakeholders and technical advisors to review the purpose of the assessment and the constructs to be measured, along with available research findings to determine which accommodations allow for valid inferences” (p. 8).

Guideline 4-D. Provide information and support to ensure that accommodations are implemented in an appropriate manner.

Burns, E. (1998). *Test accommodations for students with disabilities*. Springfield, IL: Charles C. Thomas.

In this book, Burns described the legal, practical, and theoretical implications of various types of accommodations related to format (print size, timing issues), response modes, and scoring. For example, his discussion of time extensions included consideration of motivation and fatigue as well as validity issues.

CTB McGraw-Hill. (2005). *Guidelines for inclusive test administration*. Monterey, CA: Author.

These guidelines were meant to increase the valid interpretation of individual test scores for students with disabilities and English language learners. The guidelines specifically

stated, “Agencies using both individual student results and summary results must have some awareness of the specific conditions of testing. At the individual student level, agencies must interpret the results appropriately, necessitating a specific awareness of the testing conditions. To facilitate appropriate interpretation of individual student results, testing accommodations decisions and use should be well documented” (p. 10).

Laitusis, C. C., & Cook, L. L. (Eds.) (2007). *Large-scale assessment and accommodations: What works?* Arlington, VA: Council for Exceptional Children.

This book covers a variety of issues on testing accommodations for students with disabilities and includes three sections on policy, research and development, and practice. Of particular relevance to this guideline are the introduction (Elliott), Chapter 9 (Crawford and Tindal), and Chapter 13 (Ewing), which provide guidance to both policymakers and practitioners about making decisions of appropriate test accommodations and the implementation of these accommodations. In addition, research findings on the inconsistent use of IEP assigned testing accommodations by older students (Ewing) and recommendations for future implementation (Elliott) are particularly important.

Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research and Practice, 16*, 174-181.

This article introduced a data-based approach as a way to help teachers make decisions about testing accommodations for individual students with learning disabilities. This research provided insights into the accuracy of teacher judgments in terms of which accommodations resulted in improved performance for specific students. The article concluded with a list of recommendations for practitioners.

Koretz, D., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22*(3), 255-272.

This article provided an example of how one state (Kentucky) conducted research to ensure that accommodations are implemented in an appropriate manner. The state was attempting to accomplish the simultaneous goals of including students with disabilities in the statewide assessment while also taking into account the consideration of the validity of test scores. Koretz and Hamilton analyzed the levels of inclusion, the kinds of assessment accommodations offered, the performance of students with disabilities, and the relationships between performance and the use of accommodations on both multiple-choice and open-response formats. The study findings showed that while most students were included in the assessment, some of the scores obtained may not be deemed as reliable since accommodations had been used inappropriately. The authors discussed other research and policy implications of their findings on state policy.

Shriner, J. G., & Ganguly, R. (2007). Assessment and accommodation issues under the No Child Left Behind Act and the Individuals with Disabilities Education Improvement Act: Information for IEP teams. *Assessment for Effective Intervention*, 32, 231-243.

This article provided a brief synthesis of research on the impact of testing accommodations as well as information on several studies on making decisions about testing accommodations on state accountability assessments for students with high incidence disabilities. The authors pointed out the need for more monitoring of testing accommodations to determine which are used routinely and how helpful they are. The article included a form that can be used by practitioners for monitoring students' accommodation use. The authors also provided areas for future research with a particular focus on the need for more studies that evaluate the role and effect of social-behavioral accommodations that aim to motivate students and encourage them to stay on task.

Guideline 4-E. Adjust the reading assessment approach to address the needs of some groups of students that cannot be met by typical test design or accommodation procedures.

Allman, C. B. (2004). *Test access: Making tests accessible for students with visual impairments: A guide for test publishers, test developers, and state assessment personnel* (2nd ed.). Louisville, KY: American Printing House for the Blind.

This document described a set of guidelines for making tests accessible to students with visual impairments. The author called the guidelines a “work in progress” and promised to routinely update and revise the document as additional information was collected and research results were learned. The document presented a good discussion of how computers can be used to provide accommodations tailored to a student's needs as long as the construct remains relatively unaffected. The document also included a discussion of the need to alert test users when a braille version of a test has been rescaled as a result of dropping items. The guidelines emphasized that producers of computer-based, large print, or braille formats needed to work with test publishers to verify that the test material had not been altered in content or purpose and that any change in test format maintained test validity.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Several standards within Chapter 10 of the collaboratively developed standards for educational and psychological tests focused specifically on how test developers should keep in mind the needs of test takers with disabilities and provide testing accommodations. The most pertinent standards are 10.1 and 10.10. Standard 10.1 cautioned test developers, test users, and administrators to take steps to ensure that the

test score inferences accurately reflect the intended construct rather than any disability. Standard 10.10 cautioned test professionals to consider available information about each test taker's experiences, characteristics, and capabilities when considering test modifications. To summarize, the cited standards imply that test developers, test users, and administrators need to take individual characteristics into account when assessing students with disabilities.

CTB McGraw-Hill. (2005). *Guidelines for inclusive test administration*. Monterey, CA: Author.

These guidelines provided a framework for classifying test accommodations particularly Category 3 accommodations, which are likely to change what the test measures and consequently affect the interpretation of an individual's scores. The guidelines recommended carefully considering the relationship among the test content, the desired inference, and the accommodation when interpreting individual student scores obtained using Category 3 accommodations.

Thurlow, M. L., Johnstone, C., Thompson, S. J., & Case, B. J. (2008). Using universal design research and perspectives to increase the validity of scores on large-scale assessments. In R. C. Johnson & R. E. Mitchell (Eds.), *Testing deaf students in an age of accountability* (pp. 63-75). Washington, DC: Gallaudet University Press.

This chapter addressed the application of universal design of assessment principles to large scale assessments, focusing particularly on the considerations and implications for students who are deaf or hard of hearing. In addressing ways to avoid introducing bias into the assessment, the authors provided a set of questions that can be used to address those situations where a disability (e.g., deafness) precluded the performance of skills that a test is to measure (e.g., match the sounds of words). The authors provided examples of how the reading assessment approach can be adjusted, such as allowing for the replacement of the skill with an alternative skill or allowing an accommodation not typically considered for the assessment.

Principle 5: Reporting of reading assessment results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results.

Guideline 5-A. Provide clear and concise score reports that are appropriate for a diverse audience.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

The *Standards for Educational and Psychological Testing* address all aspects of testing. The standard most relevant to providing clear language to all audiences when reporting scores is Standard 5.10. The Standard states: "When test score information is released to

students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used” (p. 65).

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Tech. Rep. No. 326). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

This technical report examined how states report results on assessments and offered guidelines for best practices, both in content and format. The authors used research from cognitive psychology, communications, and education, basing current practices data on reviews of assessment reports from 1984 and 1989. Guidelines were developed based on literature review and research, as follows: (a) know the audience and the purpose, (b) keep it simple, (c) be clear, accurate, comprehensive, and balanced, (d) use techniques to direct readers’ attention, and (e) use a format that suits the purpose. The report also included a user-friendly checklist for effective score reporting.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17, 145-220.

This article describes a study that investigated the ways in which reporting student scores on standardized tests and guides are accomplished in 11 states and selected provinces in Canada and by two test publishers. The report outlined evidence on the concerns for how results are reported and understood by different audiences and included literature on score reporting and relevant professional guidelines. Samples of actual score reporting documents are included in the text. The authors provided a comprehensive discussion on findings, including suggestions on approaches that appear to improve the readability of score reports. Among the suggestions are the use of headings and other devices to organize reports, highlighted sections, graphical displays, specifically designed reports for different audiences, and personalized score reports. The report also outlined problematic features of score reports, which included lack of information about the purpose of the reports and on precision of test scores, statistical jargon, and specific design features that are not conducive to readability.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Tech. Rep. No. 430). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Teaching.

In this study, researchers interviewed 59 policymakers and educators to investigate the readability of the NAEP Executive Summary Reports. Researchers found readers sometimes misunderstood the reports and had difficulty understanding the presentation

of results. Readers often misunderstood or simply ignored technical language (such as statistical terms) in the reports. The authors made several recommendations for improving the readability of the reports: (a) use simple, readable charts, tables and figures, (b) ensure that charts are readable without text and are field-tested, (c) reduce the amount of statistical jargon, (d) minimize the amount of technical data and the number of technical discussions, (e) include a glossary, (f) provide an introduction to scales, and (g) produce reports for different audiences. Overall, the authors recommended focusing the reports on the most important information and keeping them short. In general, this article emphasized the need for making score reports readable and clear.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

For test developers, the overarching principle in the “Reporting and Interpreting Test Results” section of the *Code of Fair Testing Practices in Education* is to report scores accurately and provide interpretation information. Specifically, Guideline C-7 stated that test developers should “...provide test results in a timely fashion and in a manner that is understood by the test taker” (p. 8). The *Code* also acknowledged the need for test users to interpret results accurately and clearly.

Trout, D. L., & Hyde, E. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

This study used focus groups of 257 participants to determine, among other things, what student score-reporting information is most useful for users or stakeholders. Each user group (teachers, principals, district administrators, parents, etc.) identified different aspects of assessment results as important. Parents, in particular, expressed the need to understand scores. The authors stated that parents “...do not understand scaled scores, prefer scores they understand such as raw scores or percentages, wanted definitions and descriptions of scoring terminology, expressed concerns about barriers in language and limited access to Internet, and want a Spanish report to be available” (p.16). By understanding the needs, usage patterns, and the different types of information wanted by the different score report user groups, reports can be made more understandable, easy-to-use, and actionable for each stakeholder group.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36, 301–335.

This study examined how well five altered data displays communicated NAEP information by interviewing educational policymakers. The purpose of the study was to see to what extent the changes in designs aided in understanding the information. According to the study, the designs produced accurate and quicker responses, suggesting an improvement in understandability. The study adds to the need for clear and concise

ways to communicate data for score reporting.

Guideline 5-B. Pilot score reports with all relevant groups of score users.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 145-220.

The purpose of this article was to survey current approaches to reporting student-level results on large scale assessments. The authors reviewed legislative requirements for reporting individual student results on statewide assessments, summarized professional and technical standards relevant to reporting scores on state assessments as well as literature related to score reporting, and examined a number of score reports from state assessments and made recommendations for improvement. In particular, the authors recommended that “Reports should be piloted with members of the intended audience and...consideration should be given to the creation of specially designed reports that cater to the particular needs of different users” (p. 208). The authors also stated, “Research should investigate potential differences among members of different demographic groups with respect to the interpretation of assessment results and should identify ways to effectively communicate assessment results across these groups” (p. 211).

Guideline 5-C. Detailed information about the assessment and score results is available in a document that is accessible to all test takers and score users.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Among the standards for educational and psychological tests comprising this document, the standard most relevant to providing detailed information about the reporting of assessment and score results is Standard 13.14. The Standard states, “In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores” (p. 148).

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 145-220.

The authors described in detail different formats of interpretive guides accompanying score reports for large scale assessments. Best practices for information presented in these guides included (a) answers to general questions about the assessment such as the purpose of the test, content, and items on the assessment; (b) the format of the assessment; and (c) where parents could obtain more information. The content of the assessment was provided in the guides, as well as suggestions for improving performance. The authors also described the score reports in detail and provided

definitions for technical terms. The authors suggested piloting score reports with members of the intended audience and creating specially designed reports for different audiences.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice*, 10, 16–18.

This study investigated how well 279 teachers understood the interpretive information of student score reports for a state testing program. Teachers were given a questionnaire that included a 17-item test and a hypothetical score report. Half the teachers were given interpretive score information. Some areas on the score report were misunderstood by many teachers even when they were given the interpretive information. Those without interpretive information scored the lowest on the questionnaire. This article demonstrated that score users need additional information to explain and help interpret scores and that this information needs to be understandable by those users.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

The guidelines for test developers in the “Reporting and Interpreting Test Results” section (Section C) of the *Code of Fair Testing Practices* are relevant here. They include Guideline C-1, “Provide information to support recommended interpretations of the results” and Guideline C-2, “Provide guidance regarding the interpretations of results for tests administered with modifications” (p. 8). Guideline C-6 for test users is also relevant: “State the intended interpretation of test results for groups of test takers” (p. 9).

Thurlow, M. L., Quenemoen, R. F., Lazarus, S. S., Moen, R. E., Johnstone, C. J., Liu, K. K., Christensen, L. L., Albus, D. A., & Altman, J. (2008). *A principled approach to accountability assessments for students with disabilities* (Synthesis Rep. No. 70). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

The authors of this report indicate that an important characteristic of a principled approach to inclusive systems of assessment for accountability is, “Reports are provided to educators, parents, students, policymakers, community members, and the media, with a clear explanation of results and implications (p. 14). They say that test administrators “...have a responsibility to ensure that data are used in ways that are consistent with the purpose of each assessment” and to see that “reports are readily available and accessible, and include cautions about misinterpretation of data” (p. 14).

Guideline 5-D. Provide information regarding the precision of reported scores for all relevant groups.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational*

and psychological testing. Washington, DC: American Psychological Association.

The importance of providing adequate information to interpret test scores is clearly stated in the background section of the validity chapter of this document containing standards for educational and psychological tests: “Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use” (p. 9). Standard 2.11 states, “If there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard error of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended” (p. 34). Standard 10.7 states, “When sample sizes permit, the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities should be investigated and reported by the agency or publisher that makes the modification. Such investigations should examine the effects of modifications made for people with various disabilities on resulting scores, as well as the effects of administering standard unmodified tests to them” (p. 107). The implications of these standards is that in order to support valid inferences, it is important to provide information about the reliability and standard error of measurement for scores for students with disabilities obtained under standard and accommodated or modified conditions.

Educational Testing Service. *ETS standards for quality and fairness*. (2002). Princeton, NJ: Author.

The purpose of *ETS Standards for Quality and Fairness* is to provide a benchmark of excellence for all ETS products and services. Standard 11.16 stated, “Present score information or other assessment results about population subgroups in a way that encourages correct interpretation and use” (p. 56). Standard 8.6 states, “Evaluate the reliability and standard error of measurement of reported assessment results for studied population groups, if the need for such studies is indicated and if it is feasible to obtain adequate data” (p. 43).

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

One of the key points made in the *Code of Fair Testing Practices in Education* is Guideline A-9 for test developers: “Obtain and provide evidence on the performance of test takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed” (p. 4).

National Association of School Psychologists. (2002). *Large scale assessments and high stakes decisions: Facts, cautions and guidelines*. Retrieved May 14, 2008, from http://www.nasponline.org/resources/factsheets/highstakes_fs.aspx

This document highlighted the factors influencing large-scale assessment, summarized precautions to take in implementing high stakes testing programs, and offered basic guidelines to policymakers and administrators. The authors urged states to distribute information about the amount of error in test scores and cautioned educators and parents about the limitations of tests. The information provided implies that adequate interpretation of test scores for students with disabilities requires some estimate of error in the test scores.