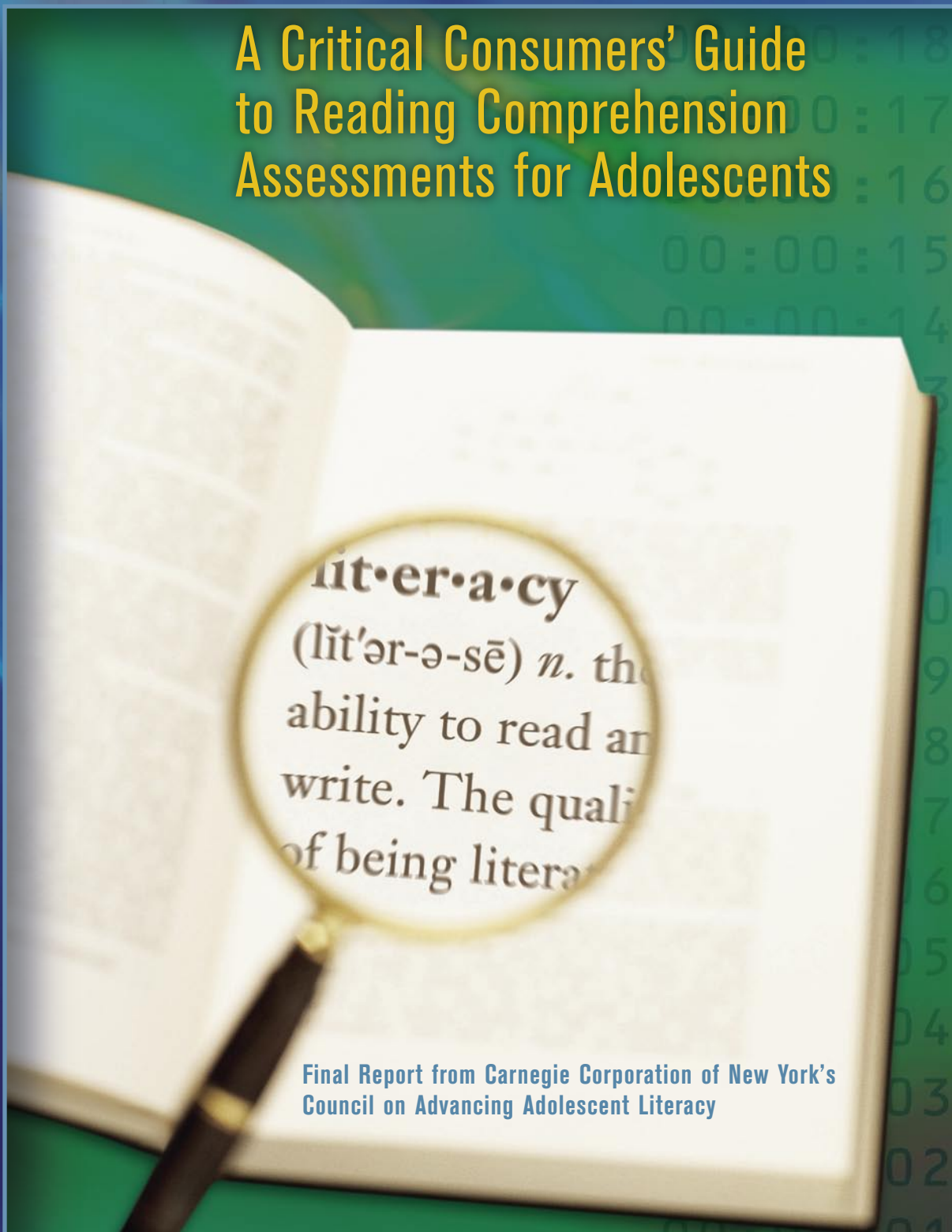
Measure for Measure

A Critical Consumers' Guide to Reading Comprehension Assessments for Adolescents



Final Report from Carnegie Corporation of New York's Council on Advancing Adolescent Literacy

Leila Morsy
Harvard Graduate School of Education

Michael Kieffer
Teachers College, Columbia University

Catherine Snow
Harvard Graduate School of Education

Carnegie Corporation's Advancing Literacy program is dedicated to the issues of adolescent literacy and research, policy, and practice that focus on the reading and writing competencies of middle and high school students. Advancing Literacy reports and other publications are designed to encourage local and national discussion, explore promising ideas and incubate models of practice, but do not necessarily represent the recommendations of the Corporation. For more information, visit: www.carnegie.org/literacy.

Published by: Carnegie Corporation of New York.

Suggested citation: Morsy, L., Kieffer, M., Snow, C.E. (2010). *Measure for measure: A critical consumers' guide to reading comprehension assessments for adolescents.* New York, NY: Carnegie Corporation of New York.

Measure for Measure

A Critical Consumers' Guide to Reading Comprehension Assessments for Adolescents

Leila Morsy,
Harvard Graduate School of Education

Michael Kieffer,
Teachers College, Columbia University

Catherine Snow,
Harvard Graduate School of Education

Final Report from Carnegie Corporation of New York's Council on Advancing Adolescent Literacy

Council Members

Carnegie Corporation of New York

437 Madison Avenue New York, NY

Council on Advancing Adolescent Literacy (CAAL)

Chair, CAAL

Catherine Snow
Patricia Albjerg Graham Professor of Education
Harvard Graduate School of Education
Cambridge, MA

Council Members

Mary Laura Bragg
Former Director, Just Read! Florida
Tallahassee, FL

Donald D. Deshler
Director, Center for Research on Learning
The University of Kansas
Lawrence, KS

Michael L. Kamil
Professor, School of Education
Stanford University
Stanford, CA

Carol D. Lee
Professor of Education and Social Sciences
Northwestern University
School of Education and Social Policy
Learning Sciences
Evanston, IL

Henry M. Levin
William Heard Kilpatrick Professor of Economics and Education and Director, National Center for the Study of Privatization in Education
Teachers College, Columbia University
New York, NY

Elizabeth Birr Moje
Arthur F. Thurnau Professor, School of Education; Faculty Associate, Research Center for Group Dynamics, ISR; Faculty Affiliate, Latina/o Studies
University of Michigan
Ann Arbor, MI

Mel Riddile
Associate Director for High School Services
National Association of Secondary School Principals
Reston, VA

Melissa Roderick
Hermon Dunlap Smith Professor, School of Social Service Administration
University of Chicago
Chicago, IL

Robert Schwartz
Academic Dean and Professor of Practice
Harvard Graduate School of Education
Cambridge, MA

Council Coordinators

Gina Biancarosa
Assistant Professor, School of Education
University of Oregon
Eugene, OR

Michael Kieffer
Assistant Professor
Teachers College, Columbia University
New York, NY

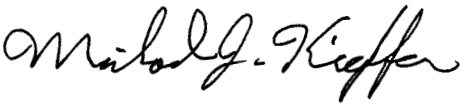
Signatories

Gina Biancarosa

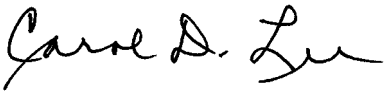
Mary Laura Bragg

Donald D. Deshler

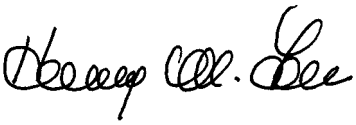
Michael L. Kamil



Michael J. Kieffer



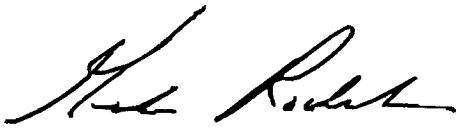
Carol D. Lee



Henry M. Levin

Elizabeth Birr Moje

Mel Riddile



Melissa Roderick

Robert Schwartz

Catherine Snow

Introduction

Although millions of dollars and weeks of instructional time are spent nationally on testing students, educators often have little information on how to choose appropriate assessments of adolescent reading for informing instruction. This guide is designed to meet that need, by drawing together evidence about nine of the most commonly-used, commercially-available reading comprehension assessments and providing a critical view into the strengths and weaknesses of each. In so doing, we focus on the utility of assessments for the purposes of screening groups of students to identify those who struggle and diagnosing the specific needs of students who struggle. Motivated primarily by the many questions that we receive from principals, literacy coaches, and district curriculum leaders about diagnostic assessment for students in grades four through twelve, this guide aims to provide those decision-makers with the tools they need to make informed decisions.

Why Would We Want More Testing?

With the increasing pressure place on schools by the standards and accountability systems, few educators can avoid the increasingly central role that tests play in schools. Given these mandates, readers might ask why we would advocate for more testing in middle and high schools. The simple reason is that the standards-based tests used in accountability systems are designed to serve a single purpose—to demonstrate how many students in a school or district have met performance standards. As a result, these tests are not adequate for the purposes of informing instruction. Although they may be able to tell educators *who* struggles with reading, they cannot provide insight into *why* these students struggle. As a result, educators must be equipped with other screening and diagnostic tools for identifying students who struggle in particular areas of reading and for placing students into appropriate interventions.

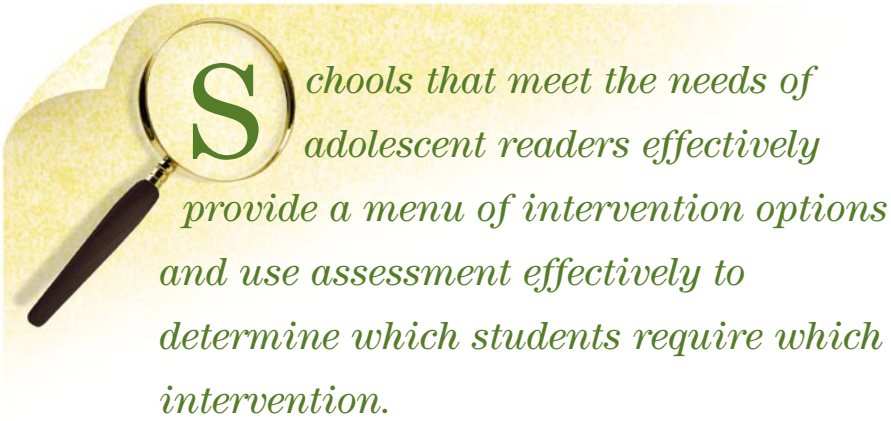
Researchers and educators have long known that effective diagnosis of students' skills and difficulties is fundamental to the successful teaching of reading. One of the great advances in early reading instruction has been the effective use of assessments to diagnose students' instructional needs, to identify students who need extra help, and to monitor their progress over time—in effective first grade classrooms, it is almost universally accepted that you test before you teach. However, teachers in grades four through twelve too often lack the tools and practices to gain instructionally useful information about their students' strengths and needs in reading.

Becoming Critical Consumers

There is no shortage of commercially-available tests, many of which sell themselves as ideal assessments for informing instruction, yet tests are not all created equal and no single test can meet all purposes. Educators must critically examine these assessments to determine how they define reading comprehension, how they assess comprehension of a variety of types of text, and what information they provide about students' strengths and difficulties. The best screening assessments will be efficient and accurate measures of students' ability to comprehend the range of texts that students are required to read at a particular grade level. They will provide teachers with a shallow amount of information about all students, with which they can judge not only the overall abilities of their group of learners but also identify those students in need of intervention or enrichment.

For this subset of learners, diagnostic assessments then provide deeper information to help educators judge which skills to target and what the nature of instruction should be. The most helpful reading

diagnostics not only measure how far behind a student is, but also identify the componential skills with which a student is struggling. Often, students differ a great deal in their profiles of strengths and weaknesses; while one student may have substantial difficulties in reading simple words aloud, another might have well-developed word reading skills but very limited vocabulary knowledge. Lumping students with dramatically different instructional profiles into a single reading intervention program will fail to meet their needs. When a reading diagnostic informs teachers more precisely about where their students' reading skills are breaking down, interventions can be better targeted to meet the reading needs of individual learners. As described in Deshler, Palinscar, Biancarosa, & Nair (2007), schools that meet the needs of adolescent readers effectively provide a menu of intervention options and use assessment effectively to determine which students require which intervention.



An early review of available screening and diagnostic assessments for mainstream students in the middle grades showed us that information was remarkably scant and scattered. This guide thus seeks to pull together relevant information about several widely-used and commercially-available assessments, combined with close analysis of the tests themselves, with the aim of informing the decisions of teachers and curriculum leaders. In so doing, we do not suggest that this list is exhaustive; rather we hope that this guide provides educators with insight into the assessments they are most likely to see as well as a range of issues to consider when they encounter assessments not covered here. Each section in this guide is about a single assessment and includes the information we considered most relevant to the selection of screening and diagnostic assessments for students in grade four through twelve.

What do Reading Comprehension Tests Test?

All reading assessments suppose an underlying "construct" of reading comprehension, that is a conceptualization of the skills and knowledge that comprise the ability to make meaning of text. This construct then is "operationalized" or made measurable through the selection of passages, the writing of questions, and (in the case of multiple-choice question) the creation of distracting incorrect answers. Thus, understanding the validity of information provided from a test of reading comprehension must start with an understanding of what the construct of reading comprehension actually is.

One group of reading researchers defined reading comprehension as: "the simultaneous construction and extraction of meaning through interaction and involvement with written language" (RAND Reading Study Group, 2002; p. 11). This definition recognizes that reading comprehension is a process involving an active interaction between the information provided by the text and the information, experiences, and actions provided by the reader—a reader who can recite statements of the meanings within the text without integrating them with her knowledge has not comprehended the text, but neither has a reader who hallucinates a fantasy of a narrative without reference to the written text.

Moreover, researchers agree that reading comprehension is multi-dimensional (i.e., made up of many different skills and practices) and developmental (i.e., changes over time). In addressing the multi-dimensionality of reading comprehension, the RAND Reading Study Group describes three sources of variation that interact to make the comprehension process more or less challenging—reader characteristics, text characteristics, and characteristics of the activity of reading. Each of these sources of variation is also situated within a socio-cultural context in which they can be understood[1].

Readers Differ

First, there is variation in *reading characteristics*, including the skills, knowledge, and preferences that students bring to the task of reading. Students differ in their vocabulary and linguistic knowledge, skills at reading words accurately and fluently, knowledge of specific content, strategic abilities to attack texts, and motivation to read on their own, all of which contribute to their ability to comprehend texts. Individual differences on these various dimensions are often greater for adolescents than for students just learning to read (e.g., Biancarosa et al., 2006; Buly & Valencia, 2002; Hock et al., 2006).

For the most part, these are the differences that assessments are designed to measure. However, tests differ in the number and range of skills that they measure; while some tests provide a single reading comprehension score, other assessments provide scores that cover a range of component skills. None of the tests we reviewed cover the entire range of skills and characteristics that reading researchers have identified as important, but several of them provide information about a number of skills central to reading comprehension, most notably vocabulary.

Texts Differ

Another source of variation that often interacts with students' characteristics is variation in *text characteristics*, including the vocabulary, language structures, text structure, genre, and background knowledge assumed in the text that students read. Although educators sometimes speak of students' "reading level," a given student might be quite successful at reading a text that is in a familiar format and about a favorite topic such as dinosaurs or Justin Timberlake, but then struggle to read an academic text about a novel concept, even if it is at the same level as measured by a readability formula. Indeed, researchers examining adolescents' literacy practices outside of school have demonstrated that many students considered "struggling" based on academic assessments can demonstrate high-level comprehension of sophisticated texts they select in other contexts (e.g., Moje, 2000).

Thus, although the focus of screening and diagnostic assessments is capturing individual differences among readers, the picture of these abilities will depend to a large degree on the texts readers are asked to read. In our review, we found that the texts used to test comprehension vary a great deal across assessments; while some of the assessments described in this guide include a range of fictional narratives, literary nonfiction (such as biographies), and expository texts, others include only a single type of text. Moreover, while some assessments provide an overall reading level for students, other assessments provide sub-scores representing students' abilities to read different types of texts. These differences between assessments are not inconsequential, but rather have a major impact on the information available about students' abilities to learn from the range of texts they encounter in middle and high school classrooms.

Reading Activities Differ

A third source of variation is in the *activity* that defines the reading task itself. Over the course of a school day, students may be asked to read a Website to learn about a new concept, study a textbook to memorize a sequence of historical events, read a math problem to find a solution, or read a poem to analyze the poet's use of figurative language. As students enter middle and high school, the range of reading activities broadens dramatically as reading becomes less a subject to be learned and more the medium through which content knowledge is learned.

Thus, the specific activity through which students demonstrate their comprehension in a given test matters a great deal. In our review, we have found that these activities can differ considerably across tests; whereas one test might ask students to identify the main idea of a passage in a multiple-choice question after reading, another test might ask students to complete sentences within the reading passage itself, and a third test might ask students to retell a story they have just read aloud. In critically analyzing tests, educators must examine how the activities in the assessments relate to their expectations for their students' comprehension.

Reading Changes over Time

It is also important to acknowledge that reading comprehension is a developmental process that looks far different in eighth grade than it does in second

grade. As students develop in their cognitive skills and knowledge of reading, expectations about what it means to comprehend rise. As students enter middle and high school, they are increasingly asked to think about what they have read in complex ways; for instance, they are asked to learn new concepts from reading, to apply what they learn from reading to new situations, and to synthesize information across different texts they have read. At the same time, texts become more sophisticated in the complexity of ideas presented and in the language in which they are

presented; sentences become longer, more rare words appear, abstract language replaces concrete objects, and connections between ideas are made more implicit. In addition to becoming more difficult, texts become more specialized and differentiated, as the basal readers of elementary school are replaced with subject-specific texts in middle and high school. These latter texts often include content knowledge, text structures, and vocabulary particular to the content-area in which they are embedded, and engage readers in a diverse range of activities, from extracting conclusions from charts in social studies to solving math problems to interpreting a play for philosophical meanings[2].

The developmental nature of reading means that diagnosing the reading comprehension ability of adolescents is more challenging than diagnosing reading comprehension among third graders. In particular, assessments should not only capture the increased sophistication of the reading task in the middle and high school years, but should also capture the specialization of the many tasks that comprise reading comprehension for older readers. Educators must think carefully not only about what the assessments they use consider "grade-level" text but also how those assessments capture or fail to capture the processes involved in reading in different content-area classes. This is one area in which reading comprehension assessment may have not yet caught up with the work of researchers or the needs of educators, but it is nonetheless worth considering.

What Should an Assessment System Look Like?

Do's

Given the complexities of reading comprehension described above, it is clear that assessing comprehension is not a simple task. Moreover, our review of commonly used assessments confirms what many have suggested—no single test can serve all purposes. Rather, a rational and purpose-driven system of assessment is required. Snow (2003) describes several requirements for such a system, including the capacity to identify individual children as poor comprehenders (i.e., screening), the capacity to identify subtypes of poor comprehenders for the purposes of differentiating instruction (i.e., diagnosis). She also indicates that an assessment system must reflect the authentic outcomes in reading that educators believe are important—that is, educators who seek to promote critical thinking in their middle school students cannot use sixth grade assessments that emphasize only factual recall. Such a system would include off-the-shelf standardized assessments, but would certainly not end with them. Rather, it would emphasize the ongoing formal and informal assessment (ranging from end-of-unit exams to writing conferences) that skilled teachers use on daily basis to guide their instruction[3].

Such a system would provide relevant and up-to-date information for teachers at strategically useful times. Middle schools might focus initially on

screening all entering sixth graders and high schools on all entering ninth graders, to get a picture of the average achievement level of the new cohort as well as identify students in need of more intensive diagnostic assessment. For this reason, we conducted our analyses of comprehension tests on the sixth grade version. That said, teachers at each grade level will likely need information about their students' skills as well as ongoing monitoring of their progress as students move through the grades.

Creating a coherent system of assessment requires that educators at various levels (including district curriculum leaders, school literacy coaches, and classroom teachers) work together to ensure that they have assessments to meet each of their specific purposes (e.g., screening, diagnosis, monitoring progress) and that educators can interpret the results of these assessments in systematic ways that can inform instructional decisions. Readers who want to learn more about creating such systems can look to Boudett, City, & Murnane (2005) for more advice, as well as Deshler, Palinscar, Biancarosa, & Nair (2007) for insight into how assessment data fits into a larger instructional plan for adolescent struggling readers.

It is also worth noting that educators should consider carefully the many different costs involved in creating an assessment system. In creating this guide, we were not able to estimate the costs in using the tests reviewed, in large part because the costs of using an assessment well go far beyond the cost of the test materials themselves. When creating a strategy for assessment, districts and schools must also consider the costs of training teachers to administer the assessments, costs of supplemental personnel to provide individual administration for some tests, costs in instructional time that is taken up by assessment, and costs in time and professional development that will support teachers to use the data to inform their instruction. Not only do the costs of the materials themselves vary considerably across tests, but also these related costs vary considerably across types of tests.

Don'ts

In the current frenzy of testing, it is worth acknowledging the real costs of excessive assessment without critical consideration of its utility. Undoubtedly, educators reading this document will be able to identify their own examples of instances when schools and districts have used assessments without a clear sense of purpose, used assessments for purposes other than those for which they were designed, or used assessments without a purpose in mind at all. Here we highlight *a few examples of inappropriate uses of assessments* that we have seen in various districts in which we have worked or conducted research.

1) **Using last year's state standards test scores to decide which students require which interventions.** Under pressure to align instruction with the content assessed on state standards tests, some schools have decided that these tests should be the only basis on which instructional decisions should be made. Unfortunately, while these tests may tell educators *who* to teach, they rarely provide information about *what* to teach, especially in the case of reading comprehension. Students may fail to meet reading standards for a variety of reasons, and additional assessments will be necessary before schools can determine which interventions are appropriate for which students.
2) **Using tests designed to look as similar as possible to the state standards tests to decide which students require which interventions.** As with the practice above, this practice involves districts investing resources and time to create "interim" assessments that indicate how well students will do on the end-of-year state tests. Although these might provide some general sense of students' progress toward the standards, educators must consider whether that information is worth the cost in instructional time and district resources. Because these are designed to be as similar as possible to the summative end-of-year tests, these tests will provide only limited information about what skills to teach or about which interventions to provide to which students.
3) **Providing teachers with a wide range of assessment choices without training or direction about which to use for what purposes.** As should become clear throughout this report, choosing which assessments to use for what purposes is not a simple or

straight-forward task. Moreover, administering and analyzing the data provided by reading assessments can require substantial teacher expertise. When districts provide teachers with boxes of assessments but little training on how to use them, teachers are likely to waste instructional time on testing without gaining useable information about students' skills and needs.

4) **Providing teachers with no standardized screening or diagnostic assessments in the interest of furthering their use of informal, formative techniques.** Although teachers should be supported and encouraged to use a range of informal formative assessments to drive their daily instructional decisions, they should also be provided with tools with which they can gain measures of students' skills that are objective, reliable, and comparable across classrooms and schools.

Avoiding these inappropriate uses of assessments and meeting the requirements above for creating a rational system of assessment requires thoughtful and reflective leadership on the part of the district and school officials in charge.

About this Report

Commissioned by the Carnegie Council on Advancing Adolescent Literacy, this report is designed to be a critical consumers' guide to several of the most commonly used commercially available assessments. Based on input from the Council, we pursued a process of selecting assessments that educators in middle and high schools would be most likely to encounter and critically reviewing those assessments for their utility for the purposes of screening and diagnosis. In so doing, we sought both to collect objective data about each assessment and to make subjective judgments about how to characterize the assessments relative to each other. The authors of this report attempted to check our subjective judgments against those of all the co-authors as well as those of thoughtful practitioners who reviewed this report. However, ultimately the judgments expressed herein should be considered the informed opinions of the authors rather than empirically-validated truths. We provide details on the selection and analysis of tests in the appendix.

What We Learned About the Tests

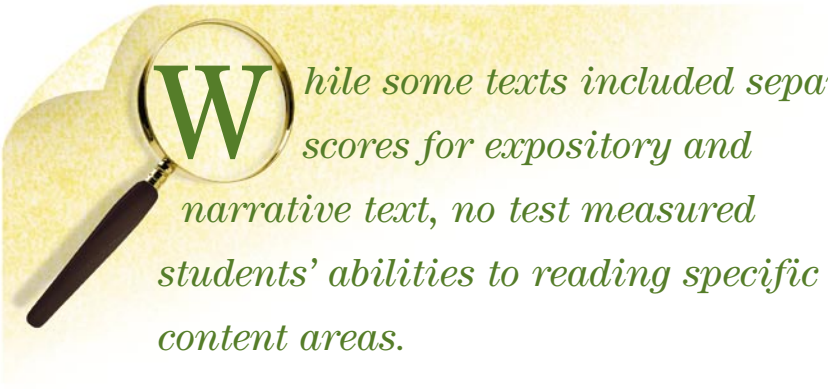
Table 1 presents our ratings for each assessment on the dimensions we identified as most important. For each category, we rated the assessments as high, medium, or low based on the information provided and none, if the feature listed was not included at all (e.g., when the assessment did not include poetry as a type of text). These ratings are relative to the utility of currently available assessments, to the extent that we considered the range of assessments on each dimension, rather than comparing them to some "ideal" assessments. A rating of high indicates that we could confidently recommend the assessment for the specific purpose (e.g., assessing that type of reader or particular skill), a rating of medium indicates that we have reservations about recommending the assessment for that purpose, and a rating of low indicates that we could not necessarily recommend the assessment for that purpose. We encourage readers to use this table to compare tests and select tests appropriate for specific purposes.

In addition to illuminating the strengths and weaknesses of different assessments, these ratings illustrate several trends that were common across all tests. Four trends are most prominent:

1) **Most of the assessments examined emphasized inferential questions of some type.** A majority of the comprehension questions on most of the tests required students to not only extract literal information from the text but also make an inference of some kind. Examples of inferential questions are displayed in Table 2.

2) **None of the tests examined emphasized critical thinking tasks.** Although many of the tests included challenging questions, the difficulty of these questions often resulted from the sophistication of their language or the attention to subtleties required to make appropriate inferences than from the level of critical thinking required. Examples of what we would consider critical thinking tasks at the sixth grade level include synthesizing knowledge across texts, critiquing an author's point of view, or composing an essay in response to literature. This was somewhat surprising, given that such tasks appear in many national and state standards; for instance, the example of a critical thinking question in Table 2 is based on items from the California Standards Test for sixth grade.

3) **Tests varied in the extent to which they included content-area passages, but no tests targeted content-specific reading skills or knowledge.** While some texts included separate scores for expository and narrative text, no tests measured students' abilities to read in specific content areas. When content-area texts did appear, they were often balanced with texts from other content areas such that the total score would be an average across skills with different texts. This is somewhat disappointing given the widespread acknowledgement that reading proficiency in different content areas requires different skills and knowledge, as described above.



While some texts included separate scores for expository and narrative text, no test measured students' abilities to reading specific content areas.

4) **Tests varied on a continuum between screening and diagnostic functions, and there was usually a trade-off between efficiency and information about individual differences.** Some tests (e.g., the SRI, the DRP, and the ITBS) were reasonably well-suited to be used as efficient whole-group screening assessments to identify which students struggled; these typically required less time and teacher expertise. Other tests (e.g., the QRI, DAR) were better suited to be used as diagnostics for providing richer information about individual differences in the componential skills involved in reading comprehension; these tests typically involved much more time, were individually administered, and often required more teacher expertise. However, we also found that several tests (e.g., SDRT, GRADE, Gates-MacGinitie) could be considered hybrids, providing more information than a single reading comprehension test score, but requiring less time than individually administered diagnostic tests. These assessments cannot necessarily be used alone to diagnose student needs (especially for those who struggled to decode words) but do provide information about a somewhat wider range of skills such as vocabulary, listening comprehension, and specific levels of comprehension.

In the pages that follow, we provide individual reviews of nine commonly used and commercially available assessments for adolescent readers. Each review includes a brief description of the design of the assessment, its operational definition of reading comprehension, and its strengths and weaknesses, based on our close analysis of the sixth grade version of the assessment. We also provide Website information for each assessment for readers to use to research them in more depth, investigate the costs of the materials, and order the assessments. We invite readers to use these reviews in combination with Table 1 as a reference when making decisions about which assessments to use for specific purposes. At the same time, we encourage readers to be critical consumers of our judgments as well, to check our opinions against their own experience and thoughtful consideration of the content and purposes of the assessments they choose. This may be particularly important for readers concerned with students at the high school level; although many of our overall statements about the tests are true at all grade levels, there may be some differences in the emphasis of skills for the ninth or tenth grade versions compared to the sixth grade version we examined. We hope that this report provides some answers about the tests currently available, but we also hope that it raises some important questions for readers to consider when deciding how to best use assessment to discover—and meet—the needs of adolescent students.

TABLE No. 1. *Comparative Chart of Sixth Grade Reading Comprehension Assessments, with Consensus Ratings of High, Medium, or Low on each Dimension*

Assessment	Degrees of Reading Power (DRP)	Diagnostic Assessment of Reading (DAR)	Gates-MacGinitie (GMRT)	Gray Oral Reading Test (GORT 4)	Group Reading Assessment and Diagnostic Evaluation (GRADE)	Iowa Test of Basic Skills (ITBS)	Qualitative Reading Inventory (QRI)	Scholastic Reading Inventory (SRI)	Stanford Diagnostic Reading Test (SDRT)
USEFULNESS FOR DIFFERENT PURPOSES									
Group screening to identify who struggles with comprehension	M	L	H	H	H	H	L	H	M
Diagnosing why individual students struggle	L	H	H/L*	M	M	L	H	L	M
Identifying strengths & weaknesses of a whole class	L	M	H/L*	M	M	L	L	L	H
Matching students to texts	H	L	L	L	L	L	M	H	L
Monitoring progress over time	H	L	M	M	M	M	L	H	M
READERS: Achievement Levels									
Struggling readers	M	M	M	H	M	M	H	M	H
Average-performing readers	M	H	H	M	H	H	H	H	M
Above-average performing readers	M	M	M	L	M	L	M	M	L
READERS: Component Skills									
Vocabulary	L	H	H*	L	H	H	L	L	H
Oral language comprehension	L	L	L	L	M	L	M	L	L
Background knowledge	L	L	H*	L	L	L	H	L	L
Comprehension strategy use	L	L	H*	L	L	L	H	L	M
Word reading accuracy & fluency	L	H	L	H	L	L	H	L	L
ACTIVITIES: TYPE OF COMPREHENSION QUESTIONS									
Emphasis on recalling facts	L	L	L	M	L	L	H	L	H
Emphasis on identifying the main idea	M	H	M	H	M	M	M	H	M
Emphasis on making inferences	H	H	H	H	H	H	H	H	H
Emphasis on critical analysis or synthesis	L	L	L	L	L	L	L	L	L
TEXTS: TYPE AND LENGTH									
Story	0	L	H	H**	M	L	L	H	M
Literary non-fiction	0	L	L	L	L	M	H	L	0
Poetry	0	0	L	0	0	L	0	0	0
Exposition	H	H	H	M	M	M	H	H	M
Argumentation and persuasive text	0	0	0	0	0	0	0	0	0
Document and procedural materials	0	0	0	0	0	0	0	0	M
Relative length of texts to other texts used in assessments	M	M	L	M	M	H	H	L	M
ADMINSTRATION REQUIREMENTS									
Efficiency of administration	H	L	L/H*	L	H	H	L	H	H
Ease of administration & analysis of results	H	M	L/M*	M	H	H	L	H	H

(*High only if item-analysis is conducted; **High only at lower levels)

TABLE No. 2. | *Categories of question types and examples*

Question Type	Example
Factual – Questions that require looking back into the text for directly stated evidence and identifying the literal restatement or paraphrase of the evidence in the answer choices.	▪ *"Who put up the sign?"* ▪ *"In the fields, it was difficult to gather —"* ▪ *"The passage says that country gets more snow because-"*
Inference – Questions that require students to make connections across one or more statements, or make connections between pieces of information in the text and their background knowledge. In these questions, the evidence is not directly stated.	▪ *"You can tell from the poster that —"* ▪ *"Why did Grandmother seldom go to the movies?"* ▪ *"Why does the author include the detail about the chimpanzee's behavior?"*
Main Idea – Questions in which students must identify the "gist" or central message of a passage. To answer these questions, students generally need to identify the more and less important information, making inferences across several sentences.	▪ *"What is the main idea of the passage?"* ▪ *"Which of the following is the best title for the passage?"*
Critical Thinking – Questions that require higher order thinking skills such as synthesizing information across texts, analyzing an author's point of view, or evaluating evidence for a claim.	▪ *"Which of the following sources would provide the best evidence for the author's position in this editorial?"*
Cloze – Questions in which a word has been removed from a sentence and students need to use understanding of the context of the surrounding sentences to fill in the blank; cloze sentences can draw primarily on sentence-level inferring skills or can require students to have a representation of the passage as a whole.	▪ *"The streets were _____."*
Question that can be answered without reading the text – These questions draw exclusively on background knowledge, including specialized knowledge of topics in English-language arts. Although they provide little to no information about students' comprehension of the texts in the assessment, they could potentially be informative about aspects knowledge that influence comprehension.	▪ *"What kind of house did Abraham Lincoln grow up in?"* ▪ *"What is the difference between a critique and review?"* ▪ *"Which of the following is not an opinion?"*

Note: Due to copyright protections on the assessments reviewed, the examples above (as well as those in the reviews below) are not taken from the tests themselves but are illustrative cases designed by the authors to be as parallel as possible in content and format as those in the tests.

Degrees of Reading Power

Critical Description

The Degrees of Reading Power (DRP) test is a group-administered assessment that may be used to determine students' overall reading level for the purposes of selecting texts or identifying students who are substantially below grade level. On the DRP, students read expository texts of increasing length and difficulty and choose a word from four choices to complete a cloze sentence (i.e., a sentence with a missing word) embedded in the texts. Completing these cloze sentences always draws on students' sentence-level understanding, often draws on students' understanding of the sentence before and after the cloze sentence, and sometimes draws on their passage-level understanding[4]. The DRP provides a single measure of overall reading comprehension ability, scaled on the DRP scale, on which several texts have also been scaled for difficulty. As a relatively short assessment with two test forms and a vertically equated scale, it can be used several times during the school year to measure students' growth in overall reading ability.

The strength of the DRP lies in its ability to provide teachers with students' reading levels, match students to texts easily, and measure growth over time. Because the test includes texts of a wide range of difficulty levels, it can be used with students of diverse ability levels and may be most useful as a beginning-of-the-year assessment for teachers who do not know their students' reading levels. The DRP usefully provides two scores—an independent reading level and an instructional reading level, to indicate to teachers which texts the student can read with 90 percent comprehension and 75 percent comprehension, respectively. The DRP scores for students and for leveling text difficulty are on the same scale, which makes student scores relatively easy to interpret—a independent level score of 60 means the student can read texts as difficult as 60 on the DRP text difficulty scale without guidance from a teacher.

However, the DRP cannot provide fine-grained information about students' reading comprehension levels nor does it provide information about the sources of reading comprehension difficulties. For instance, it will not provide insight into students' decoding and fluency skills, vocabulary knowledge, or ability to make inferences. In addition, the DRP scale is not widespread and teachers might have a hard time finding texts that are already leveled on the DRP scale. Although the collection of books that have been leveled on the DRP scale include texts of a variety of genres, the DRP assessment itself includes only expository texts and thus may not necessarily be a basis for valid inferences about students reading level with other types of texts. When compared to the assessment to which it has the closest resemblance, the SRI, the DRP has a less commonly used scale and employs cloze questions that draw more heavily on sentence-level understanding (as opposed to text-level understanding). Compared to other screening assessments, it has the disadvantage of not providing comprehension-related subscores such as reading vocabulary.

The design of the DRP implies a unidimensional definition of reading comprehension in which students' ability to complete a sentence within a text represents their overall comprehension of that text. It does not consider vocabulary as a separate construct, although the increasing difficulty of the texts is explicitly related to the increasingly density of low-frequency words as well as many other factors, including increasing word length, sentence length, and passage length (as well as other factors not explicitly taken into account in the readability formula used such as increasingly demands on background knowledge). The DRP is not written to provide information on content-area-specific reading skills, although it includes expository text exclusively. In summary, educators can look to the DRP as an adequate measure for determining students' overall reading level, but should compare it to similar measures for screening purposes and be sure to combine it with more diagnostic measures for students who struggle with comprehension.

TABLE No. 3. | *Characteristics of Degrees of Reading Power by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ Designed to measure how well students process and understand increasingly difficult prose text. ▪ Designed to measure skills used during reading rather than summarizing or analyzing after reading.
What is it actually measuring?	▪ Basic surface comprehension of prose text of increasingly difficult readability levels. ▪ Measuring whether students self-monitor their comprehension.
Overall strengths	▪ Not time-consuming. ▪ Useful for leveling students – the readability formula provides a simple way for teachers to gauge the difficulty students experience when reading texts of certain levels. ▪ Providing a readability index and test scores that are on the same scale, for ease of comparison. ▪ Scale allows for measuring growth over time.
Overall weaknesses	▪ Measures overall comprehension but does not give details about the breakdown of comprehension. ▪ Only a limited number of texts have an assigned DRP readability index. ▪ Readability formula does not work for non-linear texts, such as poetry.
For what kind of reader will the assessment give the most information?	▪ The DRP is not designed with high- or low-level readers in mind. It is appropriate for leveling most children on a single dimension of reading comprehension since passages range between relatively easy to quite difficult. ▪ The test will not differentiate difficulties that relate to decoding or fluency as opposed to comprehension. ▪ Little information will be captured about readers who have not developed decoding and fluency skills.
What are subset score categories within each subtest?	▪ None.
Administration	▪ Group administered. ▪ The test is not timed but the DRP Handbook estimates most students will take about 45 minutes to complete the assessment.
Texts	
Number of texts	▪ 10
Types of texts	▪ Expository.
Will specific background knowledge help a student answer certain questions?	▪ Although the texts tap a range of background knowledge skills, little specific background knowledge will put a student at an advantage and help them answer the questions.
What kind of content knowledge (including ELA) is required?	▪ Little ELA content knowledge is required.
Readability formula	▪ DRP readability scale (range 0-100) based on the Bormuth formula. The readability scale is only informational for well-written prose texts (no grammatical errors, well-organized).
Items	
What kinds of multiple-choice questions are included?	▪ Only cloze passage multiple-choice questions. ▪ Because there are no different question formats, once a student 'gets' how the question works, there is little room for measurement error due to question format.
Questions that can be answered without reading the text	▪ None.

TABLE No. 3.	*Characteristics of Degrees of Reading Power by Key Categories (continued)*
In the comprehension section, how many factual questions are included?	▪ No explicitly factual questions.
What does a typical factual question look like? How straightforward is the evidence?	▪ n/a
In the comprehension section, how many inferential questions are included?	▪ All questions require some level of inference.
What kinds of inferential questions are included?	▪ All cloze questions. ▪ Many of the questions require anaphoric inferring (i.e.: the ability to infer what a pronoun such as "I" or "she" refers to). ▪ In the easier passages, the sentences are grammatically simple. ▪ Successfully completing typical DPR cloze questions involves looking back into the text at the 2 or 3 sentences leading up to the cloze word. Students never need to look ahead into the passage in order to understand which cloze word fits best. ▪ All questions require students to self-monitor their understanding. ▪ No questions require students to predict, make explicit connections between ideas, summarize, or directly use background knowledge.
How many main idea questions are included?	▪ Most cloze questions require students to understand two or three sentences and sometimes an entire paragraph. No question requires understanding the entire passage.
What makes them difficult?	▪ n/a
What kinds of question stems?	▪ There are no questions stems; only four multiple-choice answers.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is not considered a construct separate from reading comprehension or assessed directly in the DRP. It is indirectly assessed through increasingly difficult vocabulary in the passages. ▪ Students do not get a separate vocabulary score.
Is difficult vocabulary necessary to answer the questions?	▪ The vocabulary in the passages is increasingly difficult. ▪ The vocabulary in the answer choices are generally common words, although they often have multiple meanings that come into play in selecting the correct answer.
Statistics	
Reported psychometric qualities	▪ Reliability: • K-R 20=.95. ▪ Validity (criterion-related): • Readability of passages correlated with difficulty of items (r=.95), • Correlations reported for other (unspecified) comprehension tests (r=.75 and .85).
Norming sample	▪ Year: 1999. ▪ Size: n=48,000. ▪ Location by weighted percentages: Midwest (16.4%), Northeast (31%), South (21%), West (31.7%). ▪ Diversity by weighted percentages: Native American/Alaskan Native (2.6%), Black (22.8%), Hispanic (10.2%), White (60.7%), Pacific Islander/Asian (3.7%).
Contact Website	http://www.questarai.com/products/drpprogram

Diagnostic Assessment of Reading (DAR)

Critical Description

The DAR is an individually-administered diagnostic assessment designed to identify students' strengths and weaknesses on a range of reading-related skills. These include overall comprehension of basic expository texts, oral reading fluency and accuracy, word reading accuracy, spelling, and oral vocabulary. The assessment is adaptive, that is teachers determine where to begin oral reading and silent reading comprehension by first administering the word recognition subtest. Each grade level has one passage, followed by four reading comprehension questions that require students to recognize accurate inferences and an opportunity for students to retell what they have just read.

The DAR would be most useful as a diagnostic assessment to identify the weaknesses of a small number of students who struggle with comprehension, especially those who struggle with grade-level expository texts. The strengths of the assessment lie in its ability to help teachers determine whether lack of comprehension stems from poor word-reading skills, poor fluency, or weak vocabulary skills. The individual administration of the DAR allows teachers to ask clarifying questions of students when answers are ambiguous or unclear, and gives teachers a greater opportunity to understand the strengths and weaknesses of individual students. Compared to other individually-administered diagnostic assessments such as the QRI, the DAR is relatively brief to administer and has the added benefit of targeting oral vocabulary, which was measured in no other assessment studied.

However, the DAR has several weaknesses compared to other assessments. Although it assesses a range of comprehension-related skills, it does not provide deep information about comprehension itself. With only a single passage and four questions per grade level, it provides limited information about students' specific difficulties with comprehension. For example, it would not be helpful in determining whether a student's struggles lie in their inability to extract factual information from the text or an inability to make inferences based on that information. It will also provide more limited information about students' comprehension skills when they are reading passages above or below their word reading level. In development of the test, the test-makers report only moderate correlations with the Gates MacGinitie Reading Test, perhaps because the latter is a grade-specific test with many more texts per grade level. In addition, although the oral retell has the advantage of providing teachers with insight into students' ability to orally summarize a text, this portion of the assessment relies heavily on teachers' judgment of students' responses, and thus makes it difficult to compare scores across classrooms or schools.

The design of the DAR implies that reading is a multi-dimensional construct comprised of a range of component skills. Reading comprehension is operationalized as the ability to provide the "gist" of the passage sufficiently to provide an oral summary and identify correctly made inferences on a multiple choice question. The DAR is not built to measure content-area reading skills explicitly, although expository text is emphasized and texts with social studies and science topics are included. It does not measure how much students rely on content knowledge to understand texts, but students who know about certain content-area topics could be at an advantage. The test-makers measure vocabulary as a construct distinct from reading comprehension in a test that includes mostly medium-frequency, high-utility words. The vocabulary subtest also includes a small number of content-area words in the higher-level vocabulary word list (though these are likely too few to draw reliable inferences about content-area vocabulary knowledge). In summary, educators can look to the DAR to determine students' abilities on a wide range of component skills, but would be wise to combine it with another measure of reading comprehension.

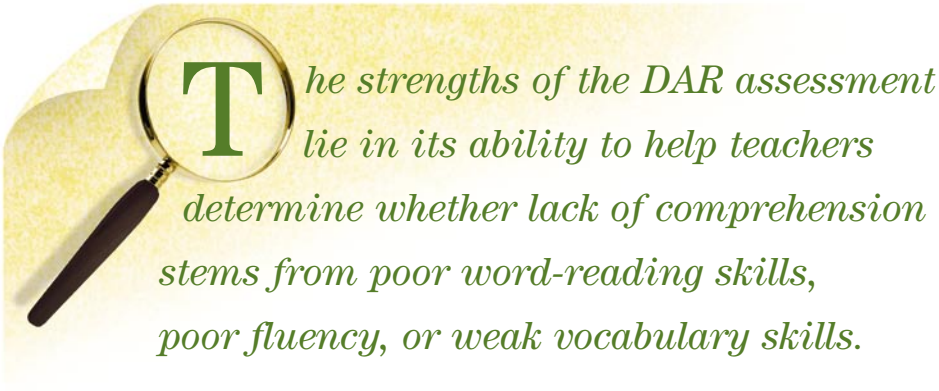


TABLE No. 4. | *Characteristics of Diagnostic Assessment of Reading by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "To assess students' relative strengths in various areas of reading and language". ▪ "To discover the areas of reading and language in which students need further instruction". ▪ "To demonstrate to students what they already know about reading and the next steps they need for improvement".
What is it actually measuring?	▪ Silent reading comprehension: whether students can answer inferential multiple choice questions and retell basic information about short, non-fiction passages. ▪ A range of other component skills such as oral reading comprehension, fluency, and word decoding, and word recognition.
Overall strengths	▪ Measures word recognition, oral reading (accuracy and fluency), spelling, and word meaning as well as silent reading comprehension. ▪ Allows teachers to ask clarifying questions to students if answers are unclear. ▪ Allows for measuring growth over time.
Overall weaknesses	▪ Does not provide information about critical thinking skills, evaluating texts on an aesthetic basis, appreciation of text, or comparing different texts. ▪ Students are diagnosed on a narrow scope of text types and only one text per grade level. ▪ Individual administration may be time-consuming. ▪ Some evidence that comprehension scores do not correlate highly with other comprehension measures.
For what kind of reader will the assessment give the most information?	▪ Because the DAR includes one non-fiction passage per grade level, it will provides some information about the comprehension skills of a wide range of students but will not provide as fine-grained information as grade-specific tests for students reading close to grade level. ▪ The DAR's many subset scores will provide the most information about students who struggle with accurate or fluent word reading.
What are subset score categories within each subtest?	▪ There are no subset score categories within reading comprehension, however, there are several other sub-scores for component skills including word recognition, oral reading, spelling, and word meaning (i.e., oral vocabulary).
Administration	▪ Individually Administered. ▪ No set limits although the test makers write that it should take approximately 40 minutes per student for the entire test.
Texts	
Number of texts	▪ 10 (across grades K-12).
Types of texts	▪ By types of text: • Stories (1), • Expository (8), • Literary non-fiction (1).
Will specific background knowledge help a student answer certain questions?	▪ Some passages cover fairly common curricular topics. It is not unlikely that this would put certain students who have read about or studied the topics at an advantage.
What kind of content knowledge (including ELA) is required?	▪ Subject of texts is relatively common and likely to appear in social studies curricula. The likelihood of a student knowing about the text topics seems relatively likely. ▪ Some background knowledge in social studies may help students answer questions without reading or referring to the text.
Readability formula	▪ None provided; texts are of varying reading levels.

TABLE No.4. | *Characteristics of Diagnostic Assessment of Reading by Key Categories (continued)*

Items	
What kinds of multiple-choice questions are included?	▪ Almost all questions (31/32) are inferential or main idea. ▪ Many inference questions require students to know specific vocabulary. ▪ Questions do not target higher order thinking skills such as synthesizing or analyzing.
Questions that can be answered without reading the text	▪ 1
In the comprehension section, how many factual questions are included?	▪ 1
How straightforward is the evidence?	▪ Evidence is straightforward and phrased almost identically to the question.
In the comprehension section, how many inferential questions are included?	▪ 31
What kinds of inferential questions are included?	▪ Inferential questions based on: • 1 sentence (9) • 2 sentences (2) • More than 2 sentences (5) • Main idea (7) • Phrase or word (8)
How many main idea questions are included?	▪ 7
What makes them difficult?	▪ Students have to synthesize information over the entire passage and based on that, draw an inference. ▪ Evidence is not directly stated. ▪ Often require specific vocabulary knowledge.
What kinds of question stems?	▪ Questions are full questions with four answer choices.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is considered a separate construct and measured in the "word meaning" section in which the teacher reads individual words and asks students to define each one orally. ▪ Measures word knowledge rather than the ability to derive word meaning from context clues. ▪ Because it is administered individually and does not include multiple choices, the Word Meaning section allows teachers to judge whether student understands the word.
Is difficult vocabulary necessary to answer the questions?	▪ 8 comprehension questions depend on students' understanding of medium- to low- frequency vocabulary, such as "seldom" or "descendant."
Statistics	
Reported psychometric qualities	▪ Reliability: split-halves correlations corrected with Spearman-Brown (Reading Comprehension): 0.96. ▪ Validity (criterion-related): • Correlations between the Gates-MacGinitie Reading Test and the DAR: vocabulary (r=.40), comprehension r=0.48), Total (r=.47).
Norming sample	▪ Year: 2004. ▪ Size: n=158 sixth graders; 1,395 students total. ▪ Location: South (30%), Northeast (48%), Midwest (8%), West (14%). ▪ Diversity: Asian-American (6%), African-American (13%), Hispanic (13%), White (61%), Other (7%).
Contact Website	http://www.riverpub.com/products/dar

Gates-MacGinitie

Critical Description

The Gates-MacGinitie Reading Comprehension test is a group-administered screening assessment composed of short passages from a relatively wide variety of genres. Each passage has a small number of associated multiple-choice questions, which draw heavily on students' abilities to make sophisticated inferences. The reading vocabulary sub-test assesses a range of grade-appropriate words by requiring students to identify a synonym for a word provided in a sentence or short phrase. The GMRT can be used in two different ways. First, with a relatively small investment of time and energy, the GMRT can serve as a screening tool to identify which students struggle with comprehension and/or reading vocabulary. Second, with substantially more time and energy, educators can analyze students' performance on individual items to determine more specifically where comprehension breakdown occurs and whether students are using unproductive strategies to understand text (as outlined in the *Manual for Scoring and Interpretation*).

The GMRT has different strengths when used in either of these ways. As a screening assessment, it is a reliable and relatively efficient test of students' overall comprehension of grade-level texts and vocabulary. Using well-established norms, educators can identify which students are performing substantially below national averages. If educators are willing to analyze students' performance on individual items, the GMRT can serve as a tool to identify group or individual strengths and weaknesses within the realm of specific comprehension processes. Test questions are built so that students' wrong answer choices provide information about productive and unproductive strategies a student is using in order to answer questions. For instance, certain questions are helpful in determining whether students are answering questions based on just an isolated word or phrase from the text while other questions are helpful in determining whether students are relying entirely on prior background knowledge rather than integrating this knowledge with information from the text. These error analyses can be used to identify trends in student performance across a classroom, and norm-referenced scores can be useful for schools or districts to use in overall planning. The reading vocabulary scores can also provide additional insight into one likely source of reading comprehension difficulty.

However, the GMRT is written for students with relatively strong fluency and phonics skills, and thus will not distinguish between word reading difficulties and difficulties in reading comprehension. If a student with poor fluency and phonics skills were to take the GMRT, their low comprehension scores would be hard to interpret and would not yield information beyond showing that the reader is struggling. As it is most often used, The GMRT provides only a single overall score for reading comprehension; breaking down and analyzing performance on individual items to understand how students' comprehension skills are weak can be time-consuming for teachers. In addition, the brevity of the GMRT passages may limit its ability to determine whether students can read extended texts such as textbook chapters or novels with understanding.

Although the GMRT provides a single comprehension score implying a unidimensional construct of reading, score interpretation using the *Manual for Scoring and Interpretation* suggests a more complex, multi-dimensional construct of reading comprehension. Although passages are short, answering questions correctly requires students to understand complex language and make relatively sophisticated inferences. Vocabulary is assessed separately and considered a separate construct from reading comprehension by the test makers. In summary, educators can look to the GMRT as an efficient and informative screening assessment to identify the sub-set of students who struggle with comprehension and also use it for more in-depth diagnostic assessment.

TABLE No. 5. | *Characteristics of Gates-MacGinitie Reading Test by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "The Gates-MacGinitie Reading Tests are designed to provide a general assessment of reading achievement... The Vocabulary Test measures the students' reading vocabulary... The Comprehension Test measures the students' ability to read and understand passages of prose and simple verse." ▪ "The objective information obtained from the tests, complemented by teachers' evaluation and other sources of information, is an important basis for: • selecting students for further individual diagnosis and special instruction; planning instructional emphasis; locating students who are ready to work with more advanced materials; making decisions about grouping students; deciding which levels of instructional materials to use with new students; evaluating the effectiveness of instructional programs; counseling students; reporting to parents and the community."
What is it actually measuring?	▪ Reading comprehension weaknesses including certain unproductive strategies to answer questions about prose text (such as relying on background knowledge instead of using information in the text; giving too much weight to a certain phrase or word in the text).
Overall strengths	▪ Identifying comprehension weaknesses including limited vocabulary. ▪ Identifying when students are using ineffective comprehension strategies. ▪ Strong psychometric basis for reliability and validity based on a series of systematic revisions.
Overall weaknesses	▪ Does not provide information about critical thinking skills, evaluating texts on an aesthetic basis, appreciation of text, or comparing different texts. ▪ Getting the most information about specific comprehension weaknesses, such as over-relying on background knowledge to answer questions, requires a lot of work from teachers. ▪ The Gates is only useful for students with strong fluency and decoding skills.
For what kind of reader will the assessment give the most information?	▪ The Gates-MacGinitie will be most informative for students reading close to grade level and least informative for students who have very high or very low reading skills for their grade level.
What are subset score categories within each subtest?	▪ There are no subset score categories within reading comprehension. However, the Manual for Scoring and Interpretation provides three lists of groups of questions a student will usually get wrong if that student lacks a particular comprehension skill. • Wrong answers indicating use of prior knowledge instead of using the text. These questions can show if a student is answering questions on the basis of background knowledge rather than by using information in the text. • Wrong answers indicating that students are giving undue weight to one section of the text instead of considering the overall logic of the text: students are not linking one sentence to the next and giving undue weight to certain sentences rather than the big picture. • Wrong answers indicating that the student is drawing the answer from a single word of phrase in the text rather than considering the overall text. ▪ Reading vocabulary.
Administration	▪ Timed: 55 minutes total (35 minutes for reading comprehension; 20 minutes for reading vocabulary).
Texts	
Number of texts	▪ 14
Types of texts	▪ Short passages from published works. ▪ By types of text: • Stories (5), • Expository (6), • Literary non-fiction (2), • Poetry (1).

TABLE No. 5.	*Characteristics of Gates-MacGinitie Reading Test by Key Categories (continued)*
Will specific background knowledge help a student answer certain questions?	▪ Subject of texts is very esoteric. The likelihood that a student would have knowledge of the various subjects of the text is highly unlikely. ▪ Even if a student did have some knowledge of the subject addressed in the text, it would benefit them very little in answering the questions.
What kind of content knowledge (including ELA) is required?	▪ Little ELA content knowledge is required.
Readability formula	▪ None provided; all texts are at approximately the same reading level.
Items	
What kinds of multiple-choice questions are included?	▪ Most of the multiple-choice questions (42/48 or nearly 90%) require inferential information. Only 6 questions are factual questions.
Questions that can be answered without reading the text	▪ None.
In the comprehension section, how many factual questions are included?	▪ 6
How straightforward is the evidence?	▪ Question or answer choices are usually phrased differently than the. evidence in the text.
In the comprehension section, how many inferential questions are included?	▪ 42
What kinds of inferential questions are included?	▪ Inferential questions based on: • 1 sentence (18), • 2 sentences (6), • More than 2 sentences (2), • Main idea (15), • Phrase or word (11).
How many main idea questions are included?	▪ 15
What makes them difficult?	▪ The evidence is not directly stated. ▪ Readers must synthesize information over most of the text and based on that, draw an inference.
What kinds of question stems?	▪ Question stems are either full questions or beginnings of sentences with answer choices that complete them.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is considered a separate construct and has its own section with 45 questions (enough questions to make the scores reliable for individuals as well as groups of students). ▪ Some words assessed have double meanings and require students to identify the correct meaning from context. ▪ Questions stems are either sentence fragments or short simple sentences with the vocabulary word underlined: • "A hoarse voice" • "She is prudent."

TABLE No. 5.	*Characteristics of Gates-MacGinitie Reading Test by Key Categories (continued)*
Is difficult vocabulary necessary to answer the questions?	▪ 11 questions are based on understanding a word or phrase such as nearly, seldom, straight away.
Statistics	
Reported psychometric qualities	▪ Reliability (KR-20): • Vocabulary (.91), • Comprehension (.92), • Total (.95). ▪ Validity (Criterion-related): • Correlation between the GMRT and ◦ CAT (vocabulary: r=.84; comprehension: r=.81; total: r=.87), ◦ ITBS (vocabulary: r=.76; comprehension: r=.77), ◦ CTBS (vocabulary: r=.72; comprehension: r=.79; total: r=.83).
Norming sample	▪ Year: 1987-1988. ▪ Size: n=77,413. ▪ Location by weighted percentages: New England/ Mideast (22.8%), Great Lakes/ Plains (26.2%), Southeast (23.9%), West/Far West (27.1%). ▪ Diversity (SES by weighted percentages): Low (24.4%), Low-Average (26.1%), High-Average (24.3%), High (25.2%).
Contact Website	http://www.riverpub.com/products/gmrt

Gray Oral Reading Test 4th Edition (GORT 4)

Critical Description

The GORT 4 is an individually administered reading diagnostic that measures passage reading fluency as well as surface comprehension skills for a wide range of students. Students read passages of increasing difficulty aloud and answer comprehension questions orally. The test is adaptive, in that administration begins roughly at students' grade level but includes passage that are easier or more difficult based on students' performance. The five reading comprehension questions for each passage target factual recall, basic inferring skills, and whether students can identify the main idea of the whole passage. Scoring allows teachers to measure students' rate and accuracy of reading as well as conduct a miscue analysis (e.g., to determine if readers substitute sounds or words, omit sounds or words, repeat words or phrases, or self-correct their errors). At successive levels, the texts increase in length, vocabulary level, and grammatical complexity. Because students read a number of passages, the teacher usually has the opportunity to evaluate students' fluency and comprehension skills on texts of different levels of difficulty.

The strengths of the GORT 4 lie in its ability to assess oral reading fluency in addition to comprehension, as well as its usefulness in measuring growth over time for students at a wide range of reading levels. The assessment results can be useful for leveling students and as a tool for choosing texts of approximately appropriate levels. In particular, the GORT 4 can be used to identify students who would benefit from interventions that target fluency skills. Compared with other individually administered diagnostics, the GORT 4 has a greater emphasis on fluency but a lesser emphasis on other component skills.

Thus, the GORT 4 is not as useful for identifying the needs of students who struggle with comprehension but do not show difficulties with fluency (or who show difficulties with fluency that result from causes other than word reading problems). The GORT 4 does not measure other sources of comprehension difficulty, including limited vocabulary knowledge, limited background knowledge, or difficulties with reading strategically. This is a disadvantage for teachers who are already aware of their students' approximate reading level and seek further information about their students' reading comprehension skills. The GORT 4 only includes one text per level, making it difficult for teachers to assess whether student struggle with different types of texts near their reading levels. Because there are few (5) reading comprehension questions per text and several questions can be answered on the basis of prior knowledge alone[5], it is possible that the test could over- or underestimate a child's comprehension ability. As an individually administered test, the GORT 4 can also be time-consuming.

The design of the GORT 4 implies that reading comprehension is somewhat multi-dimensional, with passage reading fluency as a central source of comprehension difficulty. Subject area reading comprehension is not explicitly tested, but the test includes several texts that could appear in science or social studies textbooks. Comprehension questions emphasize surface-level understanding and require relatively simple inferences. The designers of the GORT 4 do not explicitly consider vocabulary a separate construct. However, as texts difficulty levels increase, the vocabulary in the passages becomes much more complex and many questions rely heavily on vocabulary knowledge. Vocabulary words embedded in the higher-level texts and questions are not content-specific words, but rather medium-frequency words a reader would encounter in well-written prose including literature or newspaper articles. In summary, educators can look to the GORT 4 as a useful diagnostic, particularly for identifying students who struggle with fluency, but should consider using it with a subset of students first identified by a screening assessment and combining it with a measure of vocabulary.

TABLE No.6. | *Characteristics of Gray Oral Reading Test by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ Help identify students who are significantly below their peers in oral reading proficiency and who may profit from supplemental help. ▪ Aid in determining particular kinds of reading strengths and weaknesses in individual students. ▪ Document progress over time after a reading intervention.
What is it actually measuring?	▪ Rate, accuracy, and fluency of reading. ▪ Factual recall and simple inferring. ▪ Surface comprehension of short fiction and non-fiction texts of increasingly difficult readability levels.
Overall strengths	▪ The test can differentiate comprehension difficulties related to fluency. ▪ Useful for leveling students. ▪ Allows for measuring growth over time. ▪ Independent administration allows teachers to gain greater insight in students' reading comprehension skills.
Overall weaknesses	▪ Does not provide information about the nature of breakdown in comprehension beyond fluency. ▪ Some questions can be answered without reading the text which could lead to overestimating students' comprehension skills. ▪ Test is relatively complicated to score. ▪ The test does not differentiate difficulties related to vocabulary. ▪ Individual administration can be time-consuming.
For what kind of reader will the assessment give the most information?	▪ The test is not specifically designed for high- or low-level readers. Rather it will give a rough estimate of comprehension and fluency for a wide range of students. ▪ Test is designed to provide information for a wider range of students than most other assessments.
What are subset score categories within each subtest?	▪ Rate ▪ Accuracy ▪ Fluency ▪ Comprehension
Administration	▪ Individually administered. ▪ The test is not timed but test makers estimate administration should take between 15 and 45 minutes depending on students' reading level.
Texts	
Number of texts	▪ 14 (across grades K-12).
Types of texts	▪ Types of text: • Stories (8), • Literary non-fiction (2), • Expository (4). ▪ Stories predominate in the easier passages while literary non-fiction and expository texts are more common in the harder passages. ▪ Depending on grade level, texts vary in length from short paragraphs to several-paragraph essays.
Will specific background knowledge help a student answer certain questions?	▪ Some passages cover fairly common curricular topics. It is likely that this would put certain students at an advantage.
What kind of content knowledge (including ELA) is required?	▪ Certain questions (7) can be answered without reading the text if the student has some background knowledge in commonly taught social studies and language arts topics. ▪ Knowledge of common literature genres (legends, folk tales, tall tales, myths) may also help students answer some questions without reading the text.
Readability formula	▪ None provided; texts are of varying reading levels.

TABLE No. 6. | *Characteristics of Gray Oral Reading Test by Key Categories (continued)*

Items	
What kinds of multiple-choice questions are included?	▪ Most questions are inferential. ▪ Questions do not target higher order thinking skills such as synthesizing or analyzing.
Questions that can be answered without reading the text	▪ 7 See Keenan & Betjemann (2006) for an empirical analysis of this issue.
In the comprehension section, how many factual questions are included?	▪ 12
How straightforward is the evidence?	▪ Factual questions depend on the difficulty of the passage. In easier passages, evidence is quite straightforward and is phrased nearly identically as in the question. For intermediate passages, factual evidence is usually phrased differently than in the questions. For the more difficult passages, the factual evidence requires knowledge of difficult vocabulary words.
In the comprehension section, how many inferential questions are included?	▪ 58
What kinds of inferential questions are included?	▪ Inferential questions based on: • 1 sentence (11) • 2 sentences (5) • More than 2 sentences (15) • Main idea (22) • Phrase or word (5)
How many main idea questions are included?	▪ 22
What makes them difficult?	▪ Students have to synthesize information over the entire passage. Based on this, they must draw an inference. ▪ Evidence is not directly stated. ▪ In the more difficult passages, advanced vocabulary knowledge is required.
What kinds of question stems?	▪ Questions are either full questions or beginning of sentences with answer choices to complete them.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is not directly assessed. Certain questions rely on vocabulary knowledge. However, there is no separate vocabulary score.
Difficult vocabulary necessary to answer the questions?	▪ Five questions rely directly on vocabulary knowledge.
Statistics	
Reported psychometric qualities	▪ Reliability (test-retest, same form): • Fluency: .91-.94, • Comprehension: .78 - .85. ▪ Validity (criterion-related validity) for reading comprehension scores: • .41 (Gray Diagnostic Reading Test), • .62 (Wechsler Intelligence Scale for Children, 3rd Edition).
Norming sample	▪ Year: 1999-2000. ▪ Size: n=1677 (across age levels). ▪ Diversity: Sample more than 2/3 White. ▪ Location: 28 states in all regions.
Contact Website	http://psychcorp.pearsonassessments.com

Group Reading Assessment and Diagnostic Evaluation

Critical Description

The GRADE is a group-administered reading test that can provide some diagnostic information about students' reading comprehension, reading vocabulary, and listening comprehension skills. The passage comprehension section includes five passages of differing genres and multiple-choice questions that emphasize inferential skills. For instance, answering several questions requires students to draw inferences about what the author of a text would have agreed or disagreed with based on what they have read. The combination of several sub-tests provides more information about component skills than with other group-administered tests, although less information than some individually administered diagnostic tests and little information about decoding accuracy and fluency.

The GRADE provides more information about component skills than do most group screening tests. It is useful for ranking students on a handful of dimensions and for gaining basic information about large groups of students near the average range. Unlike the other screening assessments reviewed, the GRADE has specific subtests for sentence comprehension and listening comprehension. The sentence comprehension subtest requires students to choose a word from four choices to complete a relatively complex sentence; this test draws on students' knowledge of sentence structure as well as their knowledge of vocabulary in context. The listening comprehension subtest requires students to listen to sentences of varying complexity read aloud and to select the corresponding picture from four choices. A recent analysis of this sub-test suggested that the difficulty of the items comes in large part from students' understanding of idioms, their command of complex grammatical structures, and students' precision of vocabulary knowledge (Mancilla-Martinez, 2006). This subtest might be particularly useful for assessing the language skills of English language learners, independent of their reading ability. Another strength of the assessment lies in the passages included, which are authentic and engaging for students, when compared to texts in other tests reviewed. Because it has several different forms and a vertically equated scale, the GRADE can be used to show growth over time (at least for the overall reading score, if not for the component scores).

The GRADE is not very useful for providing fine-grained information about students' comprehension skills. It does not provide the information about the comprehension process a teacher would get from assessing a student individually. In addition, texts are of approximately the same readability level and therefore will provide more limited information for students reading substantially above or below grade level. Compared to some other screening assessments such as the ITBS and Gates-MacGinitie, the GRADE is relatively new and thus has not benefited from as many revisions and may have weaker psychometric properties. Finally, it is worth noting that although the makers of the GRADE provide software for analyzing the results of this assessment, several users we spoke with found this software difficult to use and excessively expensive.

The design of the GRADE implies that reading comprehension is a multi-dimensional skill that involves sufficient understanding of a text to discern the authors' purpose or make inferences beyond the text itself. The GRADE does not have many factual questions, but focuses instead on more complex inferring skills. The few factual questions included require students to draw facts from more than one sentence and recognize a paraphrased version of the evidence as stated in the text. The passages makes little vocabulary demands on students, however, some of the texts have more complex logical and linguistic structures (and therefore require more sophisticated inferential skills) than the passages in some of the other diagnostic tests we examined. In summary, the GRADE is a good choice for teachers who want information about several sub-skills for a large number of students, but should be combined with other measures, especially fluency and decoding accuracy (for at least a sub-set of students).

TABLE No. 7. | *Characteristics of Group Reading Assessment and Diagnostic Evaluation by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "The GRADE will tell you what students can and cannot do in" reading comprehension, vocabulary, sentence comprehension, and listening comprehension.
What is it actually measuring?	▪ Silent reading comprehension of grade-level fiction and non-fiction texts, including recalling facts, and drawing basic and more complex inferences based on information in the text. ▪ Componential skills related to reading comprehension, such as vocabulary skills, sentences comprehension, and listening skills.
Overall strengths	▪ Some questions require more than basic text comprehension and target complex inferring. ▪ Efficient for assessing groups of students on several dimensions. ▪ Texts are engaging, well-written, and similar to what students would encounter in an ELA classroom or reading for pleasure. ▪ Has several subtests assessing componential skills.
Overall weaknesses	▪ There are few reading passages within the reading comprehension section. ▪ Does not provide information about where breakdown in comprehension occurs beyond information provided in subtests (for instance, does not provide information about decoding and fluency skills or specific skills within comprehension).
For what kind of reader will the assessment give the most information?	▪ Students reading near the average grade-level range with well-developed fluency and decoding skills.
What are subset score categories within each subtest?	▪ Vocabulary ▪ Sentence comprehension ▪ Listening comprehension ▪ Passage comprehension
Administration	▪ Group-administered. ▪ Untimed, but test administrators suggest the entire test should take approximately 70 minutes with the following approximate times for subsections: Vocabulary (15 minutes), Sentence comprehension (20 minutes), Listening comprehension (10 minutes), Passage comprehension (25 minutes).
Texts	
Number of texts	▪ 6
Types of texts	▪ Short to medium length passages of slightly varying difficulty. ▪ Some excerpts are from published literature. ▪ By types of text: • Expository (2), • Literary Non-Fiction (1), • Story (3). ▪ Texts are similar to what students would encounter in a classroom or reading for pleasure. ▪ Although there is no excessively difficult vocabulary, some texts have complex syntax, sentence construction, and sequencing.
Will specific background knowledge help a student answer certain questions?	▪ Minimal specific background knowledge is required to answer the questions correctly.
What kind of content knowledge (including ELA) is required?	▪ No specific content knowledge in any subject is targeted.
Readability formula	▪ None provided. ▪ Texts are of slightly varying difficulties and approximately at 6th grade reading level.

TABLE No. 7. *Characteristics of Group Reading Assessment and Diagnostic Evaluation by Key Categories (continued)*

Items	
What kinds of multiple-choice questions are included?	▪ Most questions are inferential. ▪ Some main idea questions require students to go beyond simply knowing the main idea of a passage and require them to draw an inference about the main idea. ▪ A small number of questions are confusing and do not have a straightforward answer.
Questions that can be answered without reading the text	▪ None.
In the comprehension section, how many factual questions are included?	▪ 5
How straightforward is the evidence?	▪ The questions are phrased differently than the evidence in the text. ▪ The evidence is often over two or more sentences.
In the comprehension section, how many inferential questions are included?	▪ 25
What kinds of inferential questions are included?	▪ Inferential questions based on: • 1 sentence (3) • 2 sentences (2) • More than 2 sentences (11) • Main idea (9)
How many main idea questions are included?	▪ 9
What makes them difficult?	▪ Students need to synthesize the whole text to correctly answer the questions. ▪ In some of the main idea questions, students need to understand the main idea and drawn an inference about it.
What kinds of question stems?	▪ Full question with four multiple choice answers.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is assessed separately. ▪ Vocabulary section includes two different types of questions: • Basic word knowledge (matching words to definitions), • Understanding word meaning using context clues.
Difficult vocabulary necessary to answer the questions?	▪ There is little excessively difficult vocabulary in the comprehension passages.
Statistics	
Reported psychometric qualities	▪ Reliability (Cronbach's Alpha): • Total score: .95-.96 • Vocabulary: .84-.87 • Sentence Comprehension: .87-.88 • Passage Comprehension: .88 -.92 • Listening Comprehension: .65-.72 ▪ Validity: • Generally strong concurrent relationships with scores on ITBS, TerraNova, and Peabody Individual Achievement Test-Revised (correlations of .68 - .90).
Norming sample	▪ Year: 2000. ▪ Size: over 33,000 students at 134 sites (2,000 sixth graders). ▪ Location: Balanced across four major geographic regions in the U.S. ▪ Diversity: Ethnic, socio-economic, and community type (urban vs. rural) approximated the U.S. population data from 1998.
Contact Website	http://pearsonassess.com/haiweb/cultures/en-us/productdetail.htm?pid=GRADE

Iowa Test of Basic Skills (ITBS)

Critical Description

The Iowa Test of Basic Skills is a group-administered reading and vocabulary test appropriate for measuring whole-class comprehension levels on grade-level texts of a variety of genres and identifying students reading below grade level. The test is highly reliable and can be a good screening assessment. It will provide the most useful information for middling students who have good fluency and word reading skills, and provide some information about lower-performing students, specifically on their basic language skills such as punctuation and capitalization, spelling, and grammar. As its title indicates, it is a test of basic skills and is designed to provide insight into students' ability to answer relatively easy inferential questions about grade-level texts.

The strength of the ITBS lies in its ability to quickly provide basic information about most students. It will provide teachers with relatively straightforward information on which students are comfortable with reading grade-levels texts and recognizing the meaning of grade-level vocabulary. Compared to other screening tests, the main strength of the ITBS lies in its high psychometric qualities (reliability, evidence of relationships with other measures, and quality of the scaling) due to the extensive revisions it has undergone. The passages are relatively longer than those included in other screening tests providing more information about students' ability to draw inferences about passages that may be more similar to what they encounter in the classroom in terms of length. The ITBS also provides a section on productive language skills, including subtests on punctuation, capitalization, spelling, and usage and expression.

The limitations of the ITBS are that it only assesses students on grade-levels test and therefore will be limited in its ability to provide information about lower-performing and higher-performing readers. It does not provide any information on where the breakdown in comprehension occurs, and (as indicated by the test-makers themselves) will not be appropriate for deciding which intervention individual students should receive.

The design of the ITBS implies a construct of reading comprehension in which simple inferring, factual recall, and recognition of common vocabulary are the minimum skills a reader should master at the 6th grade level. Although several questions in the reading comprehension section measure students' ability to derive word meaning from context clues, the ITBS considers vocabulary a separate construct and provides a separate measure. In sum, educators can look to the ITBS for highly reliable information about students' overall reading comprehension and vocabulary skills, but should pair it with more diagnostic measures of other componential skills.

TABLE No. 8. | *Characteristics of Iowa Test of Basic Skills by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "The primary purpose [of the ITBS] is to provide information that can be used to improve instruction. (...) At all test levels, the ITBS has been designed to fulfill three main purposes: (1) to obtain information that can support instructional decisions made by teachers in the classroom, (2) to provide information to students and their parents for monitoring students' growth from grade to grade, and (3) to examine the yearly progress of grade groups as they pass through the school's curriculum."
What is it actually measuring?	▪ Surface comprehension of text. ▪ Silent reading comprehension: whether students can answer factual and inferential questions based on grade level texts.
Overall strengths	▪ Measures literal comprehension of published prose text similar to what students would encounter when reading for pleasure or in an ELA classroom. ▪ Efficient and reliable screening test for large numbers of average- and lower-performing students.
Overall weaknesses	▪ Does not provide information about critical thinking skills such as evaluating texts on an aesthetic basis, appreciation of text, or comparing different texts. ▪ The ITBS does not distinguish where the breakdown in comprehension occurs, beyond vocabulary knowledge.
For what kind of reader will the assessment give the most information?	▪ Average and lower-performing readers with well developed fluency and decoding skills. ▪ The test probably suffers from a ceiling and floor effect for stronger students whose comprehension skills are beyond basic and lower performing students with word reading or fluency difficulties.
What are subset score categories within each subtest?	▪ None.
Administration	▪ Group administered. ▪ Timed (only comprehension and vocabulary are considered part of the Reading section): • Comprehension: 50 minutes • Vocabulary: 15 minutes • Usage and expression: 30 minutes • Punctuation: 12 minutes • Spelling: 12 minutes
Texts	
Number of texts	▪ 7
Types of texts	▪ By types of text: • Story (2) • Literary non-fiction (2) • Poetry (1) • Expository (2) ▪ Passages from published works. ▪ Medium-length to long, grade-level passages. ▪ One text is similar to what a student might encounter in a grade-level textbook.
Will specific background knowledge help a student answer certain questions?	▪ Background knowledge about the topic is not necessary for answering the questions. ▪ Most passages refer to very common experiences. ▪ Some specialized background knowledge. ▪ Common experiences.
What kind of content knowledge (including ELA) is required?	▪ Although one passage could be found in a science textbook, no content knowledge is necessary to answer the questions.
Readability formula	▪ None provided; texts appeared to be grade-level.
Items	
What kinds of multiple-choice questions are included?	▪ Questions mostly target basic inferring skills. ▪ Questions do not target critical thinking skills such as literary analysis or synthesizing information across texts. ▪ 4 to 8 questions per passage. ▪ According to test makers the ITBS includes 8 factual understanding questions, 7 inference and interpretation questions, 6 analysis and generalization questions.

TABLE No. 8. | *Characteristics of Iowa Test of Basic Skills by Key Categories (continued)*

Questions that can be answered without reading the text	▪ 2
In the comprehension section, how many factual questions are included?	▪ 8
How straightforward is the evidence?	▪ Evidence in the text is phrased differently than the question; for some questions, evidence needs to be synthesized over two sentences.
In the comprehension section, how many inferential questions are included?	▪ 36
What kinds of inferential questions are included?	▪ Inferential questions based on: • 1 sentence (6) • 2 sentences (3) • More than two sentences (12) • Main idea (9) • Phrase or word (6)
How many main idea questions are included?	▪ 3
What makes them difficult?	▪ The evidence is not directly stated. ▪ Readers must draw an inference based on the entire texts. ▪ A few questions require students to put themselves in the character's shoes.
What kinds of question stems?	▪ All questions are complete questions with four answer choices.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is assessed separately: • High-frequency, grade-level vocabulary, • No subject matter words. ▪ 40 vocabulary words. ▪ Measures word knowledge rather than the ability to derive word meaning from context clues.
Is difficult vocabulary necessary to answer the questions?	▪ There are few difficult vocabulary words in the comprehension section.
Statistics	
Reported psychometric qualities	▪ Reliability: • KR-20: 0.899. ▪ Validity: • Five studies have examined the predictive relationship between ITBS scores and later achievement and found that 6th and 8th grade ITBS scores have high correlations with high school scores on the Iowa Test of Educational Development and the ACT (.73 to .84), moderate correlations with high school grade point averages (.38 to .61) and small to moderate correlations with college grade point averages (.21 to .45).
Norming sample	▪ Year: 2000. ▪ Size: n=20,216. ▪ Location by weighted percentages: New England and Mideast (24%), Southeast (29.3%), Great Lakes and Plains (19.8%), West and Far West (26.9%). ▪ Diversity by weighted percentages: White (70%), African American (14.6%), Hispanic (9.1%), Asian/Pacific Islander (3.2%), American Indian/ Alaskan Native (4.4%), Native Hawaiian (0.5%).

Qualitative Reading Inventory, 4 (QRI 4)

Critical Description

The QRI is an individually-administered diagnostic assessment based on a series of expository and narrative texts and designed to shed light on a wide range of sub-skills related to reading comprehension. Skills assessed include knowledge of text structure and text genre, background knowledge about specific domains, differences between oral and silent reading comprehension, and student-reported comprehension strategies, as well as oral reading fluency and accuracy. Unlike many of the other assessments reviewed, the QRI does not require a standardized method of administration, but rather is designed to be used by a knowledgeable teacher who makes a series of decisions about which passages to administer, how many passages to administer, and what constitutes high-quality responses to the questions. As a result of this non-standardized administration procedure, the QRI provides rich qualitative information about students' strengths and weaknesses, but does not provide standardized scores that allow students to be compared to national averages or benchmarks.

The QRI's strengths come from its comprehensiveness. If administered skillfully by a teacher who is knowledgeable about reading and knows her students well, it yields fine-grained information about individual students' reading strengths and weaknesses. Compared to other diagnostic tests we reviewed, the QRI will provide useful information about a wider range of skills among a very wide range of students, from students with basic decoding difficulties, to very high-level readers with the ability to think in complex ways about high-school texts. The QRI is designed for teachers who already have a basic sense of their students' reading levels and are interested in getting detailed information about specific aspects of a reader. For instance, the QRI is useful for a teacher who may want to diagnose a student on their ability to look back for information in expository texts. Among the assessments reviewed, the QRI is the only test that includes an explicit assessment of background knowledge (in the form of open-ended questions asked orally before reading). The QRI is also one of the few tests that can provide information about students' differential performance on expository texts as opposed to narrative texts, though it is important to note that some researchers have suggested that these comprehension questions for the expository texts may not focus on the information most important to reading in the content areas[6]

Although the QRI can be very informative, the downside is that many classroom teachers in middle and high schools may not have the time to learn to use the QRI or the classroom time to administer it. Compared with other assessments reviewed, the QRI will require a much more substantial investment of time in learning to use it effectively. To use the QRI effectively, teachers should first know their students' approximate reading level and invest substantial time with each student; thus the QRI cannot be used as a screening test. In addition, because the test does not provide standardized norms and is administered differently for each student, the QRI cannot be used to compare students to one another or to national averages. In addition, because the QRI relies heavily on subjective teacher judgment both in decisions about which passage to administer and in scoring of the students' responses, its results are likely to less reliable than those resulting from objectively scored multiple-choice test.

The design of the QRI implies a multi-dimensional construct comprised of a large range of componential skills. It assumes that each reader brings a different skill set, background knowledge, and metacognitive skills to their reading comprehension. Compared to the other assessments reviewed, the QRI taps the widest range of skills and activities related to reading. In summary, the QRI is an excellent diagnostic for a teacher with specific questions about specific students, with a strong knowledge of reading comprehension, and the time to learn how to use the QRI.

TABLE No. 9. | *Characteristics of Qualitative Reading Inventory by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "To estimate reading level" efficiently. ▪ "To match students to appropriate text." ▪ "To verify a suspected problem." ▪ "To determine reading level" in more depth. ▪ "To indicate growth." ▪ "To describe specific reading behaviors as a guide for intervention instruction."
What is it actually measuring?	▪ A range of reading behaviors related to comprehension, including oral reading accuracy and fluency, factual recall, main idea comprehension, and self-reported strategy use.
Overall strengths	▪ Format and prompts allow teachers greater insight into the reading processes of each student. ▪ Includes assessments of a wide range of component skills that may contribute to comprehension difficulties. ▪ Adaptive format allows test to provide information about students with a wide range of abilities.
Overall weaknesses	▪ Individual administration is time-intensive. ▪ Does not assess critical thinking, such as applying or synthesizing knowledge. ▪ Does not provide norm-referenced information about students' performance compared to peers. ▪ Requires substantial teacher training to administer in a way that yields reliable scores.
For what kind of reader will the assessment give the most information?	▪ A wide range of skill levels.
What are subset score categories within each subtest?	▪ Scores for each text read: • Prior Knowledge "Concept" questions before reading; • Acceptability (total "meaning-change" miscues); • Retell Scores: for narrative, these include setting, goal, events, & resolution. For expository, these include main ideas & details; • Comprehension questions: explicit & implicit, with and without look-backs. • Accuracy (total miscues); • Rate (correct words per minute); ▪ Independent, instructional, and frustration levels in grade levels (e.g., 5th or 6th) are provided for accuracy, acceptability, and comprehension questions.
Administration	▪ Individual administration; time varies widely based on administrator's choice in how much to assess. ▪ Assessment time could range from 30 minutes to 2 hours.
Texts	
Number of texts	▪ 7 (at sixth grade level).
Types of texts	▪ By type of text: • Literary non-fiction (3) • Expository (4)
Will specific background knowledge help a student answer certain questions?	▪ Unlike many other tests, the QRI begins with "concept questions" that assess background knowledge & allow you to gauge whether the text topic qualifies as a "familiar" or "unfamiliar" and interpret scores accordingly. ▪ Science and social studies knowledge of relatively known figures such as George Washington is helpful. ▪ Authors report that "concept question" scores correlate with students' reading comprehension scores, suggesting that they accurately represent at least some of the knowledge needed. For narrative biographies, examples of these questions are: • "Who is George Washington?" • "What does "tyranny" mean? • "Why was the Revolutionary War fought?" • "What were the results of the Revolutionary War?"
What kind of content knowledge (including ELA) is required?	▪ Draws moderately on relatively common science topics such as how temperature and humidity are related and social studies topics such as causes of the French Revolution or the importance of the Erie Canal.
Readability formula	▪ All texts are assigned grade levels based on the Dale-Chall formula, Fry Readability Graph, and Harris-Jacobson formula, each of which uses word frequency or syllables per word and sentence level. Agreement on two of the three formulae estimated the level. Pilot data confirmed that the texts were of increasing difficulty.

TABLE No. 9. | *Characteristics of Qualitative Reading Inventory by Key Categories (continued)*

Items	
What kinds of multiple-choice questions are included?	▪ Comprehension is measured through a retell (scored on the basis of propositions in the text recalled), and open-ended questions that are either "explicit" factual recall questions or "implicit" inferential questions. ▪ Think-alouds (not multiple-choice). ▪ After teacher modeling, students self-report on the strategies they use to comprehend the text, such as predicting, making connections to background knowledge, and self-monitoring for understanding.
Questions that can be answered without reading the text	▪ (8-12) Many of the implicit questions can potentially be answered based on prior knowledge.
In the comprehension section, how many factual questions are included?	▪ All of the retell can be considered factual questions, since it is scored on the basis of ideas recalled. ▪ 4 out of 8 questions for each passage are considered explicit and based on facts stated in the text.
How straightforward is the evidence?	▪ The evidence is quite straightforward, often stated directly in the text or requiring simple paraphrasing.
In the comprehension section, how many inferential questions are included?	▪ 4 out of 8 questions for each passage are "implicit."
What kinds of inferential questions are included?	▪ Questions that require students to connect one piece of information in the text with another. ▪ "Why" questions that require understanding of cause & effect based on information in the text.
How many main idea questions are included?	▪ 1 out of 8 for each passage.
What makes them difficult?	▪ Because they are open-ended questions, students need to not only grasp the gist, but put it into a brief, summarized answer.
What kinds of question stems?	▪ Open-ended questions that require single-sentence answers.
Vocabulary	
How is vocabulary assessed?	▪ Not assessed explicitly after reading. ▪ Two to three "concept" questions before reading each passage include vocabulary questions about specific content vocabulary, such as "What is a anthropologist?" "What is photosynthesis?" ▪ One question in each of two passages asked about other vocabulary, including words similar to converge or relentlessness.
Is difficult vocabulary necessary to answer the questions?	▪ In addition to the content-specific vocabulary that appears in several questions (such as above), general academic vocabulary is necessary to understand the questions.
Statistics	
Reported psychometric qualities	▪ Validity: In a small-scale study, comprehension scores on none of the 6th grade passages correlated significantly with Terra Nova scores, although prior knowledge scores did predict Terra Nova scores for a majority of passages. ▪ Reliability: For well-trained & knowledgeable scorers, inter-scorer reliability estimates were between .94 and .98. Alternate-forms reliability (based on different choices of passages) of instruction-level decisions were all above .80. ▪ Other extensive technical information is provided.
Norming sample	▪ Norm-referenced scores are not provided. The above psychometric data are based on a pilot sample of 178 students in grades 4-8 for new materials and a clinical sample of 898 students in grades 1-11, the overwhelming majority of whom were in grades 1-4.
Contact Website	http://portal.wpspublish.com

Scholastic Reading Inventory (SRI)

Critical Description

The SRI is group-administered screening assessment for identifying students' reading levels and identifying students who are reading substantially below grade level. The SRI requires students to choose words to complete cloze sentences that represent the main idea of a short text. It includes a wide range of narrative and expository texts of increasing difficulty. The SRI provides teachers with scores that are on the same scale as the Lexile Framework—a readability scale - making them easy to interpret and useful in matching students with texts that have a Lexile level.

The strengths of the SRI lie in the usefulness of the Lexile scale to match readers to texts and to show growth over time. The Lexile scale is in widespread use among book publishers; all Scholastic trade books as well as many other books come with a Lexile level, and teachers can determine the level of a text by typing in portions of it into the Lexile Framework Website. Thus, teachers can use the SRI scores to match readers to specific texts for independent reading and for use in instruction. The Lexile scores or norm-referenced scores provided by the SRI can also be used to identify students who are reading substantially below grade level. In addition, because there are several different forms of the SRI on a single vertical scale, the assessment can be used to show growth over time on one dimension of reading comprehension—namely, the ability to get the gist of short prose passages. Compared to the tests it resembles the most, the SRI has the advantages of a commonly used scale, a mixture of narrative and expository passages, and well-designed cloze questions that target the main idea of the passage.

The SRI is less useful for teachers who know their students' approximate reading levels and seek information about where individual students' breakdown in comprehension occurs. For instance, the test does not provide diagnostic information about students' vocabulary knowledge, passage reading fluency, word reading abilities, or skills in making inferences. Thus, the test will not provide teachers with specific information on particular areas of improvement in reading comprehension. Text passages in the SRI are also shorter than those found on some other assessments, suggesting that it may be limited in assessing students' ability to read extended texts, such as novels and textbook chapters, with understanding. It is also possible that a student could make progress over time in specific areas of reading comprehension that the test would not capture, such as vocabulary knowledge or fluency.

The design of the SRI implies that reading comprehension is a unidimensional construct on which students and texts can be appropriately matched. Comprehension is operationalized as the ability to complete a main idea statement at the conclusion of a short passage. The SRI does not assess content-area reading skills, although it includes a mixture of expository and narrative passages. Vocabulary is not considered as a separate construct and not assessed separately. In summary, educators should look to the SRI as a useful tool for assessing the reading level of students and matching them to texts, but should combine this assessment with a diagnostic assessment for individual readers whose specific sources of difficulty are difficult to identify.

TABLE No. 10. | *Characteristics of Scholastic Reading Inventory by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "The tests are based on a powerful new system, called the Lexile Framework that can help you accurately assess your students' comprehension levels and match them with appropriate texts for successful reading experiences."
What is it actually measuring?	▪ Students' ability to get the gist of short, narrative fiction and non-fiction passages sufficiently to correctly fill in the blank in short, summary sentences.
Overall strengths	▪ Matches students to texts using Lexile scale. ▪ Scale is useful for showing growth over time. ▪ Efficiency in identifying students who are reading significantly below grade level. ▪ Usefulness for leveling students – the readability formula (a way of determining the difficulty level of a text) provides a simple way for teachers to gauge the difficulty students experience when reading texts of various levels.
Overall weaknesses	▪ Does not provide information about skills such as analyzing, evaluating texts on an aesthetic basis, appreciating or comparing texts. ▪ Vocabulary is not assessed separately. ▪ The SRI does not distinguish where the breakdown in comprehension occurs. ▪ Length of the test; students could get tired because there are 54 different short passages.
For what kind of reader will the assessment give the most information?	▪ The SRI is not designed specifically for high- or low-level readers. The test is most reliable for students reading somewhat below grade level. ▪ The test will not differentiate difficulties that relate to fluency or phonics as opposed to comprehension. Little information will be captured about readers who have not developed basic reading skills.
What are subset score categories within each subtest?	▪ None.
Administration	▪ Can be group or individually administered. ▪ The test is not timed. ▪ The Teacher's Guide recommends setting aside 40-60 minutes.
Texts	
Number of texts	▪ 54
Types of texts	▪ By type of text: • Expository: 20 • Story: 27 • Literary non-fiction: 4 ▪ All texts are published elsewhere. ▪ Most of the easier texts are stories whereas most of the harder texts are expository. ▪ Texts are short (between ~30 and ~90 words).
Will specific background knowledge help a student answer certain questions?	▪ The questions do not pertain to the general topic of the text but rather to students' ability to synthesize the specific information in the short passages. Background knowledge about the topic of the text is not necessary to answer the questions. Furthermore, over half the passages are narrative stories containing very little school-based background knowledge.
What kind of content knowledge (including ELA) is required?	▪ Expository texts pertain to subjects such how trees get nutrients and categories of mammals. However, the sentences cannot be comprehended without referring to the organization of information in the passage. ▪ Little ELA content knowledge is required.
Readability formula	▪ The Lexile Framework (provides a common scale for matching reader ability and text difficulty); texts are of varying reading levels.
Items	
What kinds of multiple-choice questions are included?	▪ All questions are main idea. ▪ Sentences students must fill in are all summary sentences. ▪ Questions target summarizing skills. ▪ After each passage is a short statement with a missing word. Student chose which word best fills the blank by picking from four answer choices. ▪ Each passage has only one associated question. ▪ The distractors are all unambiguously incorrect based on the text. However, if a student were to not read the text, any of the four answer choices would fit grammatically in the sentence.

TABLE No. 10.	*Characteristics of Scholastic Reading Inventory by Key Categories (continued)*
Questions that can be answered without reading the text	▪ None. ▪ The statements students need to fill in are so general that without reading the text, any answer choice would fit grammatically.
In the comprehension section, how many factual questions are included?	▪ No question has its answer directly stated in the text – all questions require inferences over several important statements in the passage.
How straightforward is the evidence?	▪ n/a
In the comprehension section, how many inferential questions are included?	▪ All (54).
What kinds of inferential questions are included?	▪ Inferential questions require students to remember and synthesize the passage. ▪ Inferential questions based on: • 1 sentence (0), • 2 sentences (0), • More than 2 sentences (0), • Main idea (54). ▪ Phrase or word (0).
How many main idea questions are included?	▪ All (54).
What makes them difficult?	▪ Students must make an inference based on the entire or most of the text.
What kinds of question stems?	▪ All question stems are simple statements with a missing word.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is not assessed separately. ▪ Although some of the answer choices include relatively difficult vocabulary, it would be hard to tell from the SRI whether a student's breakdown in comprehension is linked to poor vocabulary knowledge.
Difficult vocabulary necessary to answer the questions?	▪ Some answer choices include difficult vocabulary. ▪ A small number of passages include some difficult vocabulary as well as figurative language necessary to understand the text and correctly answer the question.
Statistics	
Reported psychometric qualities	▪ Reliability: • The Lexile scale ranges from 200 to 1200. The Standard Error of Measure ranges between 55 and 83 Lexile Points, suggesting good reliability across reading levels. The most reliable scores for the 6th grade test are for students with scores between 500 and 900 (approximately 3rd grade to 6th grade level). This suggests that the test provides the most reliable information for 6th grade students reading near or below grade level. ▪ Validity: • Construct-related correlation with SDRT4: 0.91.
Norming sample	▪ Size: n=512,224. ▪ Diversity: White (66.3%), African-American (29.3%), American Indian (1.7%), Hispanic (1.2%), Asian (1.0%), and Other (0.6%). LEP (0.7%), students with disabilities (10.1%), and eligible for free or reduced-price lunch (40%).
Contact Website	http://teacher.scholastic.com/products/sri/

Stanford Diagnostic Reading Test, 4th Edition, Paper and Pencil Test

Critical Description

The Stanford Diagnostic Reading Test is a group-administered reading test that measures surface understanding of short- to medium-length texts of varying genres. The test can be considered something of a hybrid between a screening test and a diagnostic assessment. It provides information about a range of comprehension skills, vocabulary knowledge, and students' ability to scan long texts for key information, but does not provide information about fluency or word reading skills. Designed with struggling comprehenders in mind, the test provides the most information about students' abilities to answer relatively easy questions about texts that make relative low demands on vocabulary, background knowledge, and inferential processing.

The strength of the SDRT lies in its ability to provide information about particular areas of comprehension and vocabulary in which students may be struggling. It can be a relatively efficient way to gain information about large groups of low-achieving students. In addition to norm-referenced scores for comprehension, vocabulary, and scanning skills, the SDRT provides teachers with a breakdown of questions and scores by vocabulary word type, comprehension skill, and text genre. Examining these sub-scores can be useful for teachers who are seeking to identify trends in comprehension and vocabulary across groups of students as well as recognize strengths and weaknesses of individual students. Compared to other screening assessments which are typically designed with grade-specific texts and activities, the SDRT allows teachers to differentiate between students who are somewhat below grade level and those who are substantially under-performing.

The weaknesses of the SDRT lie in its limitations in providing information about higher levels of comprehension with more sophisticated texts, as well as its inability to distinguish comprehension difficulties that result from decoding difficulties. Compared to other screening assessments, it is limited in assessing students' ability to comprehend difficult grade-level texts. Compared to other individually-administered diagnostic assessments, it is limited in assessing students' oral reading fluency and accuracy. In addition, although the subscores for specific vocabulary word types, comprehension skills, and text genres can be helpful in identifying areas for intervention, these subsections have few questions and therefore low reliability. As a result, drawing conclusions about individual students' abilities in those subsections can be risky.

The design of the SDRT implies that reading comprehension is a multi-dimensional skill drawing on different processes as well as types of knowledge, especially vocabulary. As described by the test-makers, the multiple choice comprehension questions measure "practical understanding", implying that the test is geared towards evaluating whether students have reading skills adequate enough to be used for average practical purposes on a daily basis. Although a small number of questions measuring vocabulary are included in the comprehension section, vocabulary is also considered a separate construct and measured separately. In summary, the SDRT provides more information about componential skills in reading than many screening tests, but less information than many individually administered diagnostic tests; educators can look to this test as a way to gain some insight into the difficulties of low-achieving readers, but should consider combining it with more in-depth diagnostic testing for students who struggle.

TABLE No. 11. | *Characteristics of Stanford Diagnostic Reading Test by Key Categories*

Overview	
What is the stated purpose of the assessment?	▪ "The Stanford Diagnostic Reading Test is intended to diagnose students' strengths and weaknesses in the major components of the reading process. Its results can be used to challenge students who are doing well and provide special help for others who lack some of the essential reading skills. They also can be used to identify trends in the reading levels of students in the district, provide information about the effectiveness of instructional programs, measure changes that have taken place over an instructional period, and keep the community and school board informed about students' overall progress in reading."
What is it actually measuring?	▪ Surface, functional reading comprehension as used in daily life. ▪ Basic vocabulary. ▪ Scanning.
Overall strengths	▪ Many main idea questions measuring overall comprehension. ▪ Separate vocabulary assessment. ▪ Efficient screening test for large numbers of low-performing students.
Overall weaknesses	▪ Does not provide information about skills such as analyzing, evaluating texts on an aesthetic basis, appreciating or comparing texts. ▪ The SDRT 4 does not distinguish where the breakdown in comprehension occurs, beyond vocabulary knowledge.
For what kind of reader will the assessment give the most information?	▪ Lower achieving and average students who have adequate word reading skills. ▪ Because the passages and questions target basic reading skills, the test suffers to some extent from a ceiling effect for stronger readers and will not provide much information about their comprehension skills.
What are subset score categories within each subtest?	▪ Comprehension: • Recreational • Textual • Functional • Initial understanding • Interpretation • Critical analysis • Process strategies ▪ Vocabulary: • Synonyms • Classification • Word parts • Content area
Administration	▪ Group administered. ▪ Timed: • Comprehension: 50 minutes • Vocabulary: 20 minutes. ▪ Scanning: 15 minutes.
Texts	
Number of texts	▪ 9
Types of texts	▪ Genres of texts: • Story: 3 • Expository: 3 • Document or procedural materials: 3 ▪ Short- to medium-length passages. ▪ Accessible to struggling readers. ▪ No specific content-area texts. ▪ Texts are chosen to seem typical of those one would need to understand on a daily basis (instruction manual, flyer).
Will specific background knowledge help a student answer certain questions?	▪ Expository passages are about relatively common topics but the questions do not require background knowledge.
What kind of content knowledge (including ELA) is required?	▪ Hardly any specific content-area knowledge is required to successfully answer the questions.
Readability formula	▪ None provided; all texts are approximately of the same reading level.
Items	
What kinds of multiple-choice questions are included?	▪ Questions don't target higher order thinking skills such as literary analysis, or synthesizing information across texts. ▪ The multiple-choice questions assess the following: understanding words and phrases using context clues, setting, plot and sequence, cause and effect, fact and opinion, author's purpose, making predictions based on the passage, and identifying the genre of the passage and the main idea of the passage.

TABLE No. 11.	*Characteristics of Stanford Diagnostic Reading Test by Key Categories (continued)*
Questions that can be answered without reading the text	▪ 4/54 (7.5%) questions can be answered without referring to the text.
In the comprehension section, how many factual questions are included?	▪ 24
How straightforward is the evidence?	▪ For some questions, the evidence is simple and included in one sentence. For others, students need to draw evidence from two consecutive sentences. Sometimes, the evidence in the text is phrased differently than in the question.
In the comprehension section, how many inferential questions are included?	▪ 30
What kinds of inferential questions are included?	▪ The evidence is not directly stated. Students need to draw from background knowledge, put themselves in the character's shoes, make a guess using several sentences as evidence, make connections between real world experience and the text. For instance, the question will call for understanding cause and effect, although the evidence is stated in the text, the student must on their own make a causal connection between the evidence and the question, as it is not explicit. ▪ Inferential questions based on: • One sentence (7) • Two sentences (2) • More than two sentences (8) • Main idea (10) • Background Knowledge (3)
How many main idea questions are included?	▪ 8
What makes them difficult?	▪ The evidence in not directly stated. Students need to be able to read, remember, and make sense of the whole text.
What kinds of question stems?	▪ Question stems are either complete questions or beginnings of sentences with answer choices that complete them. ▪ Question stems are sometimes phrased negatively.
Vocabulary	
How is vocabulary assessed?	▪ Vocabulary is considered a separate construct and has its own section. Students must choose a word's meaning from four answer choices. No clues are included in the question stem. ▪ Measures word knowledge rather than the ability to derive word meaning from context clues.
Difficult vocabulary necessary to answer the questions?	▪ Potentially difficult words such as seldom and rarely appear in questions and change their meaning dramatically.
Statistics	
Reported psychometric qualities	▪ Reliability: • KR-20: 0.95 to 0.98. ▪ Limited validity information was available. Although test-makers showed correlations between previous versions of the test, no data was presented on correlations between the SDRT4 and other similar measures of reading comprehension.
Norming sample	▪ Year: 1990s. ▪ Size: n=33,000. ▪ Location by weighted percentages: representative of the national population based on 1990 US Census. ▪ Diversity by weighted percentages: representative of the national population based on 1990 US Census.
Contact Website	http://www.pearsonassessments.com

http://www.pearsonassessments.com

Afterword: Researchers & Practitioners Building Better Assessments

While we hope this guide illuminates many of the purposes that currently available reading comprehension assessments can serve, we also recognize that it illustrates several of the needs that current assessments cannot meet. Thus, in this section, we provide a few examples of ways in which thoughtful practitioners and literacy researchers are working to build better assessments.

Informal Assessments of Content-area Literacy

As indicated earlier in this report, none of the assessments we reviewed provide specific information about the skills of readers in particular content areas such as science, social studies, and math. One way that practitioners have worked to fill this void is to create their own informal assessments of these skills based on the classroom texts they use. For instance, content-area teachers can create their own informal reading inventories, in which they choose specific texts from their curriculum, listen to students' read these aloud, encourage students to "think aloud" describing the strategies they use to read, and asking targeted questions about their process of navigating and comprehending the text. Such inventories can be given to individual students who seem to demonstrate particular difficulties or can be modified to give to larger groups of students, for instance by having students read silently and write about their comprehension process. For more information on this and other informal techniques, readers can turn to Readance, Bean, & Baldwin (1989).

Another approach is for teachers to look closely at students' performance on content-area assessments to get a sense of their students' strengths and difficulties with the literacy demands within those assessments (and by extension within their content area). For instance, educators can assess students with specific released items from the state standards tests or National Assessment of Educational Progress and compare their performance on items that present different types of demands in reading and writing. Such analyses may not be simple or straight-forward, but with the support of teacher teams or school literacy specialists, many content-area teachers have found these activities helpful.

Supplemental Assessments of Comprehension-related Skills

As mentioned in several of the reviews above, one approach to addressing the limitations of the currently-available reading comprehension tests is to supplement them with other independent measures of the component skills in reading and language that influence reading comprehension. Although we cannot review all of these tests here, we suggest that interested readers investigate some of the tests found to be highly useful for instruction. For instance, in the area of reading fluency, the Dynamic Indicators of Basic Early Reading Skills Oral Reading Fluency test (DIBELS; Good, Kaminski, Smith, & Laimon, 2001) and the Test of Sight Word Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) are considered by many to be quick and useful tools. An alternate to these individually-administered assessments include silent word fluency tests that may be appropriate in getting an initial overview of all students' fluency; promising examples of these include the Test of Silent Word Reading Fluency (TOSWRF; Mather, Hammill, Allen, & Roberts, 2004) and the Test of Silent Contextual Reading Fluency (TOSCRF; Hammill, Widerholt, & Allen, 2006). In the area of language, the Adolescent Screening Test (ALST; Morgan & Guilford, 1984) and the Test of Adolescent and Adult Language--Fourth Edition (TOAL-4; Hammill, Brown, Larsen, & Wiederholt, 2007) are two examples of individually-administered tests that are can be used to assess the oral language skills of a sub-set of students.

Developing Assessments for English Language Learners

None of the assessments we reviewed were designed specifically with English-language learners in mind. Since many of the assessments currently available can yield valuable information about these learners, we certainly recommend that educators use whatever means they have to screen ELLs and to diagnose their strengths and weaknesses (rather than wait for more specialized tests to be developed). That said, it can be particularly difficult to diagnose the sources of reading difficulty for ELLs, because they tend to have much more diverse profiles in their vocabulary knowledge, background knowledge, and others skill related to literate language use[7].

A current research project led by David Francis and funded by the Institute of Educational Sciences aims to create a new test, the Diagnostic Assessment of Reading Comprehension (DARC), designed to test comprehension with passages that use very simple language. Francis and his colleagues have found that many ELLs who perform poorly on other standardized comprehension measures do quite well on this assessment, suggesting that these learners need instruction in English rather than comprehension itself (Francis et al, 2006).

Making Assessment More Efficient through Computer Technology

Given that we could not describe any of the assessments reviewed to be a comprehensive and complete system for screening and diagnosing reading comprehension difficulties, there is clearly a need to develop efficient and coordinated systems of assessment. In the short term, we recommend that educators piece together a battery of assessments that can serve various purposes and make strategic decisions about which assessments should be given to all students and which are better used to diagnose the skills of a sub-set of struggling students. In the long term, however, a self-contained, comprehensive assessment product will lead to more systematic and efficient data collection.

One particularly promising approach is to use computer technology to facilitate the collection and analysis of diagnostic data. Two studies funded by the Institute of Educational Sciences, one headed by John Sabatini of Educational Testing Service and another by Gloria Waters of Boston University, are working to develop computer-based assessment tools that will assess a range of language and literacy skills in a short amount of time and provide immediately useable diagnostic results. Both assessments are currently being piloted in the Strategic Educational Research Partnership's Boston Public Schools Field Site, and will likely be incorporated into a single tool called the Reading Inventory and Scholastic Evaluation (RISE). Similarly, Wireless Generation, one computer-based assessment company, is in the process of collaborating with researchers to provide technology tools for making individual assessments of reading more efficient.

Moving Forward

To support the development of new and better reading comprehension assessments, many theoretical questions will also need to be addressed. For instance, there are many complex issues involved in assessing reading comprehension in the content areas. Creating useful tests of Science literacy or Social Studies literacy will require not only thoughtful selection of typical texts from these subjects but also the careful delineation of what it means to read texts like a biologist or a historian. Another key issue is the nature of sub-types of struggling readers; although recent studies have begun to describe the diversity of skill profiles among adolescent readers, more research is needed to investigate the prevalence and nature of these profiles in different contexts, to determine whether these profiles are stable over time, and to evaluate whether students with different profiles respond differentially to interventions. As this guide has demonstrated, the currently available assessments have many strengths that educators can (and should) make use of immediately, but they also have substantial limitations that can only be remedied through a sustained research and development effort.

References

Biancarosa, G., Mancilla-Martinez, J., Kieffer, M., Christodoulou, J. & Snow, C. (2006, July) Exploring the heterogeneity of English reading comprehension difficulties among Spanish-speaking middle school students. Paper presented at the meeting of the Society for the Scientific Studies of Reading, Vancouver, Canada.

Boudett, K.P., City, E., & Murnane, R. (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning.* Cambridge, MA: Harvard University Press.

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.

Cain, K. & Oakhill, J. (2003). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology*, *76*, 697-708.

Deshler, D., Palincsar, A. S., Biancarosa, G., & Nair, M. (2007). *Informed choices for struggling adolescent readers: A research-based guide to instructional programs and practices.* Newark, DE: International Reading Association.

Francis, D.J., Lesaux, N.K., Rivera, M., Kieffer, M.J, & River, H. (2006). Research-based recommendations for instruction and academic interventions. *Practical guidelines for the education of English language learners:* Portsmouth, NH: Center on Instruction. Retrieved on August 24, 2007 from http://www.centeroninstruction.org/files/ELL1-Interventions.pdf

Good, R. H., Kaminski, R. A., Smith, S., & Laimon, D. (2001). *Dynamic Indicators of Basic Early Literacy Skills, 5th edition (DIBELS).* Eugene, OR: Institute for the Development of Educational Achievement, University of Oregon. Available at http://dibels.uoregon.edu

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (2007). *Test of Adolescent and Adult Language (4th ed.).* Austin, TX: Pro-ed.

Hammill, D.D., Widerholt, J.L., & Allen, E.A. (2006). *Test of Silent Contextual Reading Fluency.* Austin, TX: Pro-ed.

Hock, M. F., Brasseur, I. F., Deshler, D. D., Catts, H. W., Marquis, J., Stribling, J. W., & Mark, C. A. (2006). *What is the Nature of Struggling Adolescent Readers in Urban Schools?* Lawrence, KS: The University of Kansas Center for Research on Learning.

Heller, R. & Greenleaf, C. (2007). *Literacy instruction in the content areas: Getting to the heart of middle and high school improvement.* Washington, DC: Alliance for Excellent Education. Retrieved on August 26, 2007 from http://www.carnegie.org/literacy/pdf/Content_Areas_report_6-10-07_FINAL.pdf

Kamil, M. L. (2003). *Adolescents and Literacy: Reading for the 21st Century.* Washington DC: Alliance for Excellent Education.

Keenan, J. & Betjemann, R.S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, *10*, 363-380.

Mather, N., Hammill, D.D., Allen, E.A., & Roberts, R. (2004). *Test of Silent Word Reading Fluency.* Austin, TX: Pro-Ed.

Moje, E. B. (2000). To be part of the story: The literacy practices of gangsta adolescents. *Teachers College Record*, *102*, 652–690.

Morgan, D. & Guilford, A. (1984). *Adolescent Language Screening Test.* Austin, TX: Pro-Ed.

RAND Reading Study Group (2002). *Reading for Understanding: Toward an R&D Program in Reading Comprehension.* Santa Monica, CA: RAND Corporation.

Readence, J. E., Bean, T. W., & Baldwin, S. R. (1989). *Content area reading: An integrated approach* (3rd ed.). Dubuque, IA: Kendall/Hunt.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Short, D., & Fitzsimmons, S. (2007). *Double the work:* Challenges and solutions to acquiring language and academic literacy for adolescent English language learners: *A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education

Snow, C. E. (2003). Assessment of reading comprehension: Researchers and practitioners helping themselves and each other. In Sweet, A.P. & C.E. Snow (eds.), *Rethinking reading comprehension.* New York: Guilford Press.

Sweet, A. P. & Snow, C. E. (2003), *Rethinking reading comprehension.* New York: Guilford Press.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency. Austin*, TX: Pro-Ed.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco, CA: Jossey-Bass.

Wiliam, D. (2001). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.), *Curriculum and assessment*, pp. 165-181. Greenwich, CT: JAI Press.

APPENDIX

Note on Methodology

Selection of Tests

The final list of tests to review was compiled through a multi-step process. First, we examined the Websites of the thirty largest school districts in the United States to furnish a tentative list of assessments in use in grades 6 to 12 for purposes beyond summative assessment. Assessments used solely in the context of Special Education or programs to identify or classify English language learners were not included on this initial list. Although we recognize that information from district Websites may not be the most reliable representation for what districts are actually doing with these assessments, we considered it a reasonable first step for generating a list of assessments in use. Second, a handful of assessments were removed from this list, because they were designed by individual states or districts and would not be commercially available to others. Third, we excluded one assessment (the Terra Nova), because it could not be purchased by the research team, but was only available to school district officials. Finally, we reviewed the range of assessments on the dimensions we examined, and decided to add two assessments (the GORT and the GRADE) that represented unique features that had yet to be included; although these two tests were not reported as used in the thirty largest school districts, they have been used in research and clinical settings. The sixth grade student booklet, together with the administration manual and technical manual, was collected for each test. When two parallel forms were available, one form at random was chosen for analysis. In the case of adaptive tests designed for students of various ages, the entire test was examined, but analyses were focused on the portions of the test likely to be given to sixth grade students.

We chose to focus specifically on the sixth grade version of the test because of the importance of this grade as a transition between elementary school and middle school. Although some of our findings (e.g., about the number of inferential questions in each test) are somewhat specific to the sixth grade test, others (e.g., about the scores provided and the format of each test) are true across grade spans. The overall strengths and weaknesses identified for each test will be more or less accurate for tests in grades four through twelve.

Analysis of Tests

After preliminary examinations of several tests, the three authors collaborated to create a template for the elements to examine in each test. The first and second author, both experienced middle school teachers with master's degrees in literacy, then examined each assessment to complete the template, discussing disagreements as they arose. They then wrote the narrative descriptions that precede each section.

Inter-rater reliability

The ratings of high, medium, and low provided in Table 1 are based on agreement among the first author, second author, and a third reviewer, who also had a master's degree in literacy and considerable experience in assessing and instructing adolescents. Agreement in initial ratings between any two raters was quite high, approximately eighty percent across categories; in the case of disagreements, the raters re-consulted the test materials, discussed appropriate evidence, and came to a consensus. Given time-constraints and the qualitative nature of much of the descriptions, it was not possible to calculate inter-rater reliability for the remaining portions of the test characterizations.

Endnotes

[1] For more information about the nature of reading comprehension than is provided in this brief summary, readers are referred to the report of the RAND Reading Study Group (2002) as well as Sweet and Snow (2003), which is a companion piece written with a practitioner audience in mind.

[2] For more information about how the nature of reading comprehension changes and becomes more content-area-specific in the adolescent years, readers are referred to Heller & Greenleaf (2007) and Kamil (2003).

[3] For more information on formative assessment, a large and important topic that is simply beyond the scope of this report, readers are referred to Wiggins (1998), Wiliam (2001), and Shepard (2000).

[4] For a discussion of the advantages and disadvantages of the cloze format, see Cain & Oakhill (2006).

[5] For a critique of the GORT based of the inclusion of questions that can be answered on the basis of prior knowledge, see Keenan & Betjemann (2006).

[6] In her research with middle school students' science reading, Elizabeth Moje and colleagues found that the comprehension questions provided for some of the expository Science texts did not necessarily focus on the Science knowledge that content experts identified as the most important to learn from the text, and that in a few cases, the passages themselves represented superficial treatments of the science concepts.

[7] For more information on the diversity of strengths and weaknesses among ELLs, see Francis, Lesaux, Rivera, Kieffer, & Rivera (2006) and Short & Fitzgerald (2006).

437 Madison Avenue
New York, NY 10022
(212) 371-3200
www.carnegie.org