RESEARCH BASE

# mCLASS Reading 3D Atlas Book Set Technical Manual, 2nd Edition

Amplify.

# Table of Contents

# Introduction

Amplify Atlas is a leveled reading book set that was developed for use within Amplify's mCLASS®:Reading 3D™ Text Reading and Comprehension (TRC) assessment. TRC is based on an assessment approach developed by Marie Clay, author of An Observation Survey of Early Literacy Achievement (Clay, 2005). TRC is a running record (alternately known as a reading record) assessment of reading performance that allows teachers to evaluate a student's performance on the foundational skills necessary to become a fluent reader, and the ability to apply those skills to increasingly complex text.

TRC assesses reading accuracy and comprehension using a set of calibrated benchmark books. Using TRC, a teacher determines each student's instructional reading level at three benchmark administration periods during the school year and monitors student reading performance between those periods.

The Atlas book set was designed for benchmark or screening purposes to help educators make district-, school-, and student-level instructional decisions. It specifically supports teachers to provide the literacy instruction students need to be college and career ready and achieve the rigorous expectations exemplified in the Common Core State Standards for English Language Arts (CCSS for ELA; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Amplify has built on approximately ten years of experience with leveled reading assessment in TRC to develop a book set that incorporates the most up-to-date research on text leveling, reading assessment, and reading instruction to address text complexity considerations and instructional shifts called for by the CCSS for ELA. Atlas will help teachers ensure students read and comprehend increasingly complex texts so they are "college and career ready by the end of high school" (NGACBP & CCSSO, 2010, p. 3).

## Theoretical Framework

Reading proficiency is widely assessed using summative tests administered toward the middle or end of a school year. Such tests describe student ability in terms of predefined standards, reflecting progress made during the current academic year

and preparedness for the next grade level (Garrison & Ehringhaus, 2007). While once-a-year summative tests provide information about students' year- end achievements, they do not provide the time-sensitive, detailed information needed to inform instruction throughout the school year (Guskey, 2003; Shephard, 2000; Stiggins, 2002).

Black and Wiliam's review of studies on the effects of formative assessments on student achievement in reading and math across age groups and nationalities confirms that "teaching and learning must be interactive" (Black & Wiliam, 1998, p. 2). Educators require indicators that provide rich information about student progress on a continuous basis so they can adjust instruction to their students' current level of understanding. Formative literacy assessments can be administered periodically throughout the school year to provide timely information about the source of students' reading difficulties.

TRC is an elementary-level, standardized, formative reading assessment that supports educators to systematically observe and monitor student reading skills and progress and diagnose reading difficulties. TRC is based on the running record portion of Marie Clay's Observation Survey assessment (1993a, 2002), which was developed on the principles that students learn to read through teacher-guided interactions with connected text at the appropriate level of difficulty (Cox & Hopkins, 2006) and that a research-based understanding of the reading process should inform the match between students and texts (Clay, 2001). The standardization of TRC supports objectivity and allows for comparisons both between students and within an individual student over time, and the authenticity of the task makes it a valuable indicator of what processes students use during reading (Clay, 2002). Information gained from TRC helps educators determine students' reading levels and match texts to readers for a variety of tasks that promote students' literacy development. The instructional level is the level at which the text challenges students' current skills and knowledge without impacting comprehension of the text; students can read these texts independently but may need occasional support from a teacher. The independent level of text is a bit easier and gives students a chance to build confidence in their skills, learn from the text, and develop fluency; students can read these texts without the support of a teacher or parent. The frustrational level of text is too difficult for students to read for meaning by themselves and students may find reading them discouraging; students should read these texts in a shared reading environment to increase content knowledge and familiarity with written language patterns (Clay, 2002; Fountas & Pinnell, 1999).

Knowing the instructional, independent, and frustrational text levels of students in a class guides teachers in differentiating core reading instruction, identifying students who need targeted intervention (including both small-group and individualized interventions), grouping students who read at similar levels, and determining the

content of intervention programs so students receive the instruction they need to read successfully (Snow, Burns, & Griffin, 1998; Clay 2002). Experimental findings show schools that use a running record assessment like TRC to plan instruction scored higher on reading and writing achievement tests than control schools that used other types of classroom assessments (Ross, 2004), and a study of Grade 1 teachers showed the most effective teachers used running records to plan one-on-one instruction (Pressley et al., 2001). TRC also helps schools determine whether students are making adequate gains toward meeting performance targets by the end of the current school year and even the end of the next school year (Zhao & Von Secker, 2008; Von Secker, Zhao, & Powell, 2008).

TRC measures a student's reading accuracy, fluency, and comprehension. Reading words in connected text with sufficient speed and accuracy and drawing meaning from it are intrinsically linked in an interactive process in which the reader combines what is written in a text with his or her knowledge about orthography, oral language, written language, and the topic (Rumelhart, 1977) to create a mental representation of the text. This process is constrained by the limits of working memory (Kintsch & van Dijk, 1978); thus, automaticity in word identification allows the reader to focus attention on creating a mental representation of a text (Ehri, 1995; Chall, 1996; Dowhower, 1987). The process is interactive because the mental representation of a text that a reader constructs on an ongoing basis becomes another source of information for the reader to use in word identification (Rumelhart, 1977).

Reading fluency and comprehension typically develop together: more accurate readers tend to read more quickly and with greater comprehension (Fountas & Pinnell, 1996), and if accuracy declines when students encounter more difficult books, so too will fluency and comprehension (Carver, 1990). Accuracy is a necessary, though insufficient, condition for comprehension; fluency and knowledge (including linguistic and background) link these two aspects of reading development. TRC tracks growth in reading accuracy, fluency and comprehension at successively higher levels of text complexity.

## A Brief Introduction to Text Levels

Following largely from the work of Marie Clay (1991, 1993a, 1993b, 1998) and Fountas and Pinnell (1996), the text levels in TRC are situated along a text complexity gradient ranging from A to Z. These levels are rooted in both an understanding of reading behaviors at various points in the developmental trajectory (e.g., Holmes & Singer, 1961; Rumelhart, 1994; Chall, 1983; Ehri, 1991) as well as a consideration of specific text characteristics (described in the following paragraphs). Text levels are therefore ordered indicators of text characteristics and demands assigned to individual texts as a result of qualitative evaluation by and agreement between authors and/or trained reviewers.

Strictly speaking, the resulting levels possess the measurement properties of an ordinal rather than an interval scale: higher-level texts are more demanding than those at lower levels, but the difference in demands between texts at two levels may not be equivalent to the difference in demands between texts at another pair of levels (see Stevens, 1946, for further information about measurement scales). For example, the differences between a level A text and a level B text are not equal to the differences between a level R and a level S text. There are finer differences between lower-level texts than between higher-level texts because for beginning readers (typically, students in kindergarten and Grade 1), a smaller amount of progress makes a bigger difference in the types of texts that can be read than it does for more advanced readers (Fountas & Pinnell, 1999). Clay (2001) explains that reading develops through the transformation of simple processes into more complex ones rather than through an additive accumulation of skills; the steps students must take to read text at successively higher levels are necessarily larger at the beginning of reading development because more extreme transformations are required (e.g., new readers can find understanding the concept of a word and a word's representation in print challenging). Likewise, there is a larger range of variation within the higher levels compared to the lower ones (e.g., a wider variety of language patterns, genres, and topics); thus, students generally proceed through the higher levels at a slower pace (Fountas & Pinnell, 1999).

Although actual leveling procedures may differ among publishers according to proprietary protocols, they generally consider criteria similar to those provided by Fountas and Pinnell (1999, 2011): genre (e.g., fantasy, biography) and form (e.g., text formatted in chapters or sections); text structure (e.g., chronological, compare/contrast); content (e.g., quotidian activities, novel scientific concepts); themes and ideas (e.g., concrete, abstract); language and literary features (e.g., figurative language, technical language); sentence complexity (e.g., simple subject/verb constructions, embedded clauses); vocabulary (e.g., conversational language or domain-specific academic language); words (e.g., frequency, regularity of spelling); illustrations (e.g., pictures that support a story, graphics that organize information); and book and print features (e.g., length, layout). It is the interaction of these characteristics with one another and with the knowledge of the reader that makes a text more or less difficult; for instance, a new reader may find an informational text on a novel scientific concept with plenty of picture support less difficult than a familiar narrative with no picture support. Thus, think of text levels and equivalents between publishers as approximations of complexity (Fountas & Pinnell, 1999).

# TRC Measures

## Development of the Text Reading and Comprehension Measure

TRC was developed based on the running record portion of Marie Clay's Observation Survey (1993a, 2002), which uses a leveled-text gradient and focuses explicitly on reading accuracy, reading strategies, and reading comprehension. By testing a student with a series of benchmark books, each of which is an exemplar of a text level, a teacher (or another trained assessment administrator) can efficiently identify the student's instructional reading levels. Both the text reading and the comprehension portions of TRC support the assessment of student progress towards achieving many of the standards for learning included in the Common Core State Standards for English Language Arts in Reading Foundational Skills, Reading Literature, and Reading Informational Texts (NGACBP & CCSSO, 2010).

The original version of TRC was developed in 2004 by Amplify, then Wireless Generation, in collaboration with the Montgomery County Public School District in Maryland (MCPS) and Drs. Craig and Sharon Ramey of Georgetown University as part of the Assessment Program in Primary Reading (AP-PR). The goal was to develop an assessment instrument that was pedagogically balanced (addressing both word reading and comprehension) and vertically integrated (offering materials for kindergarteners through sixth graders) to provide information about all students across the reading spectrum, whether they were barely sounding out letters in second grade or reading third-grade books as a kindergartener. The initial cut points for proficiency at each grade level and time period (i.e., beginning, middle, and end of year) were established by correlating TRC performance levels to performance on external measures of reading, such as the Comprehensive Test of Basic Skills (CTBS), TerraNova Second Edition, and the Grade 3 Maryland State Assessment (Zhao & Von Secker, 2008; Von Secker, Zhao, & Powell, 2008).

Since its original development, TRC has been subject to a program of ongoing research and development. These developments include, but are not limited to, the following: the addition of benchmark book sets and associated oral and written comprehension tasks in both English and Spanish; evaluation of TRC assessment procedures and content against various state ELA content standards and the CCSS

for ELA; collection and examination of validity evidence with respect to a variety of external measures; and, lastly, specification of benchmark book set cut points that support the requirement that students are career and college ready by the end of high school.

The Atlas book set represents Amplify's continued commitment to ensuring and improving the quality of TRC as an assessment of reading performance in support of evidence-based instructional practice. Research on the Atlas book set continues beyond the documentation presented here.

## TRC Software

TRC software makes the process of administering running records more efficient and reliable than paper-based systems. Running records are widely used by reading specialists: results from a national survey of over 1,500 specialists showed that 62 percent used running records to inform instruction (Bean, Cassidy, & Grumet, 2002). While running records are traditionally administered with paper and

pencil, this method of administration requires time-consuming and complex hand-scoring and data analysis, as well as storage of a large volume of paper records that are unlikely to travel with a student from grade to grade. In addition, Marie Clay herself complained that teachers sometimes use running records in "unacceptably slipshod ways" as a result of employing nonstandard procedures (Clay 2001, p. 45). TRC software improves standardization and eliminates the labor-intensive process of traditional paper-and-pencil assessment by guiding the teacher through each step in the assessment and electronically capturing the full running record on a handheld device or computer. Historical results and details can be viewed both on the electronic device and via the web, traveling with the student from teacher to teacher and from grade to grade. Educators can review behaviors observed at any point in the history of a student's literacy development, which is critical information when working with a struggling reader.

## Description of TRC

The TRC assessment includes various components that work together to determine a student's final instructional reading level:

- Print Concepts
- Reading Behaviors
- Reading Fluency
- Reading Accuracy
- Retell and Recall

- Oral Comprehension
- Optional Written Comprehension Tasks

**Foundational Knowledge: Print Concepts and Reading Behaviors**

Prior to the assessment of connected text reading, a student's basic familiarity with book and print knowledge is assessed (i.e., before level A). The Print Concepts (PC) assessment is based on the Concepts about Print portion of the Observation Survey developed by Clay (1993a, 2002) and measures what students know about books and print. Students begin school with varying degrees of experience with text (Stanovich, 1986), and the PC assessment helps teachers identify what their students already know and what they need to be taught in preparation for learning to read (Clay, 2002).

In this assessment, the teacher and student read a short storybook together, and the teacher asks the student to demonstrate an understanding of the basic features of a book and printed text. By reading the book with the student, the teacher provides scaffolding that allows even nonreaders to express what they know about text (Vygotsky, 1978), mainly through actions (e.g., pointing) rather than verbal responses; as a result, teachers do not need to wait until formal reading instruction begins in order to gather information about their students' knowledge. Students are asked to find the cover of the book, differentiate print from pictures, distinguish upper- and lowercase letters, demonstrate that punctuation signals meaning, and recognize that strings of letters are words, among other concepts. Teachers use the TRC software to mark student responses as correct or incorrect.

When students reach the research-derived cut point on PC, they move on to the Reading Behaviors (RB) assessment (the level between PC and level A). This assessment requires a student to read a text, but rather than scoring the student's accuracy, the teacher focuses on whether the student demonstrates appropriate reading behavior patterns. Again, teachers use the TRC software to record whether the student can perform each of these behaviors, such as recognizing common sight words, reading in a left-to-right pattern, and conducting return sweeps at the end of each line. Students who reach the research-derived cut point on the RB assessment progress to reading records.

Low scores on either of these measures can serve as an early warning signal that future problems in reading are likely for the student, as they indicate the student's level of experience with text prior to schooling. This experience is foundational to a student's ideas about the nature and functions of printed text and guides the way that he or she initially interacts with print. A lack of such experience, due to either a lack of exposure or a lack of attention to print, can have far-reaching consequences if no immediate actions are taken to intensify the student's encounters with text and reading instruction in the school environment (Justice & Ezell, 2001; Adams, 1990;

Stuart, 1995; Johns, 1980). Essentially, a student who does not achieve the research-derived cut points on PC or RB assessment in kindergarten is already behind in literacy development and needs extensive opportunities to catch up with his or her peers to prevent the gap from widening (Clay, 1998).

### Text Reading and Comprehension

Once a student shows competency on the PC and RB assessments, his or her reading is assessed using leveled benchmark books. The TRC book sets include both literary and informational texts. At each text level, the student reads a benchmark book and completes a number of follow-up comprehension tasks, which may include Retell (for literary texts) or Recall (for informational texts) and Oral Comprehension Questions (OC). The student reads from a physical book while the teacher follows the interactive text on his or her device, observing and recording the student's reading errors (insertions, omissions, substitutions, or hesitations during which the teacher tells the student a word) and self-corrections by clicking on the word, categorizing the error or self-correction, and, when appropriate, writing down the word the student actually said. These data determine the student's reading accuracy, self-correction rate, and error rate. The comprehension tasks help the teacher determine whether the student understood the meaning of the text.

Determining a student's instructional text level is an iterative process that involves starting at one  level of text difficulty and working up or down from that level to a text that is read with 90–94 percent accuracy. Previous research indicates students need to read with a minimum rate of accuracy (typically 90–94%) in order to comprehend a given piece of text (Clay, 2002; Fountas & Pinnell, 1999; Fuchs, Fuchs, & Deno, 1982); thus, the instructional level determined by TRC is the level at which a student can read with 90–94 percent accuracy and show evidence of comprehension. Accuracy percentage is calculated based on the number of words the student read correctly out of the total number of words the student encountered in the text; if the student scores below 90 percent (frustrational level), a lower-level text will be subsequently presented for the student to read, and if the student scores at or above 95 percent (independent level), a higher-level text will be presented for the student to read. Once 90–94 percent accuracy is achieved, the student is asked to complete the comprehension portion of the assessment by giving a passage retell (literary texts) or recall (informational texts) and/or by answering oral comprehension questions. If a student does not pass the comprehension tasks, a lower-level text is presented if this level was not yet administered. A student's instructional level is determined as the highest text level at which a student performs with 90–94 percent accuracy and is also proficient on the comprehension section.

TRC assesses reading comprehension in various formats. The student may be asked to orally retell or recall what the passage was about, and the retell or recall is scored on a rubric ranging from 0 to 3 points, based on the number of details in the text

the student provided. Research on the assessment of reading comprehension using retell proves the method both reliable and valid (Marcotte & Hintze, 2009; Fuchs & Fuchs, 1992; Roberts, Good, & Corcoran, 2005). For some books, oral comprehension questions are available. These questions target a student's fundamental understanding and interpretation of the text. Student responses to the five oral comprehension questions are marked as correct or incorrect.

Once all of the relevant text reading and comprehension sections of TRC are administered and instructional level determined, the student's TRC reading level is compared to expert-derived cut points to determine proficiency level according to the expectations of the Common Core State Standards. Proficiency levels can be interpreted as the degree to which a student demonstrates desirable and necessary grade-level reading behaviors, indicated by the CCSS for ELA. (See the TRC Standard Setting Research Report for more information). The proficiency determinations are color-coded red, yellow, green, and blue, which correspond to the descriptors Far Below Proficient, Below Proficient, Proficient, and Above Proficient. The color-coding simplifies data analysis, so educators can easily identify students who could benefit from enrichment or intervention.

In addition, teachers can analyze the student's reading record data to identify appropriate instructional content and diagnose a student's word-reading difficulties by carrying out a qualitative analysis of student behaviors. The ratio of self-corrections to errors is informative because a high ratio indicates that the student pays attention to how the words he or she reads fit into the overall structure or meaning of the sentence or text, while a low self-correction ratio indicates that the teacher may need to work on developing this student's level of metacognition (Fountas & Pinnell, 1999). In addition, teachers can conduct an MSV analysis on the errors a student made to determine whether they happened based on Meaning cues (e.g., substituting "kitty" for "cat"), Structural cues (e.g., substituting a noun for another noun), or Visual cues (e.g., substituting a word like "cat" with a similarly spelled word like "cap"). Such MSV data provide educators with insight into the word-reading strategies a student relies upon so that the teacher can use existing strengths to expand the student's strategic repertoire (Clay, 2002).

**Optional Written Comprehension Task**

In addition to oral comprehension tasks, Written Comprehension (WC) questions may be provided starting at text levels typically appropriate for Grades 1 and 2. Reading and writing are two sides of the same coin; Marie Clay (1998) emphasizes the importance of writing in a literacy program because students can make reciprocal gains (which teachers can observe) through the use of both reading and writing activities. Although the scoring rubric does not include the quality of a student's writing in the WC assessment (scores range from 0 to 3 based on the level of understanding apparent in the student's response), the teacher can use

WC to informally observe a student's understanding of written language based on his or her written output. The WC questions used alongside the OC questions in TRC therefore provide teachers with a broader perspective on how their students process both the structure and the content of texts and respond to them in two different production modes.

Although WC items provide important instructional guidance to teachers, the WC score is not included in the determination of instructional reading level. Reading and writing are highly related, key components of literacy development; however, they are distinct, develop differently, and must be separated to ensure accurate, reliable, and valid assessment (Berninger, Abbott, Abbott, Graham, & Richards, 2002; Juel, 1988). Instructional reading level is not dependent upon writing skills, therefore, inclusion of a writing task in determination of overall instructional reading level within TRC can lead to inaccurate results and a misunderstanding of students' reading and writing skills. This can, in turn, lead to inappropriate instructional decisions. If a student's writing skills lag behind reading skills, as is commonly the case in the elementary grades, including a writing task in the calculation of the student's instructional reading level could lead to an underestimation of that student's reading ability and could lead a teacher to assign reading materials the student finds too easy.

## Description of Atlas

The Atlas book set was developed for use in screening or benchmarking assessment within Amplify's TRC assessment. This book set includes 76 books corresponding to levels A to Z of a text complexity gradient (Clay, 1991, 1993a, 1993b, 1998; Fountas & Pinnell, 1996), with CCSS for ELA-based enhancements. There are three books per level for A through U and two books per level for V through Z. Additionally, three books at level A are reserved for the assessment of early literacy skills via the Print Concepts and Reading Behaviors tasks. For more information on the Atlas book set and its development, see the Amplify Atlas Book Set Development white paper.

# Overview of Research on Atlas

Research on the Atlas book set is ongoing. This chapter describes data and analyses from completed research studies, including the purpose of each study, how participants were recruited, demographics for the participants, experimental design, and the descriptive statistics calculated for each study.

## Study A: Initial Field Study

**Purpose:** Study A was designed to examine the reliability, difficulty, and validity of the Atlas book set.

**Recruitment:** The study used two different strategies to recruit districts and schools. First, Amplify Account Management and Sales teams were asked to contact existing mCLASS:Reading 3D customers with information about the field study. This approach yielded moderate success, so another strategy was adopted in which existing mCLASS:Reading 3D customers received an informational flyer about the field study via email. Twenty-six schools were enrolled in the study using the first recruitment approach, while three schools were enrolled in the study using the second approach.

**Participants:** This field study was conducted during the 2013–2014 middle-of-year (MOY; December through February) benchmark administration period. In total, 653 students in kindergarten through Grade 5 were assessed by 47 educators in 29 schools across seven school districts.

**Demographic Information:** Participants in this field study included educators and students from across the United States, including the following geographic divisions: East North Central, Pacific, South Atlantic, and West North Central (US Census Bureau, n.d.). The sample was composed of participants from the following demographic categories: 44 percent male, 46 percent female, and 10 percent unspecified gender; 26 percent white, 12 percent black, 30 percent Hispanic, and 32 percent other race or unspecified race. The comparison of field study sample, TRC users in mCLASS nationally, and all public schools in the United States (U.S. Department of Education, 2012) is provided in Appendix 1. There are differences in the geographic location distributions among national public schools, TRC users, and the field study sample. However, the effect sizes of the differences are very small (≤ 0.01).

**Table 1. Sample Size, Demographics, and TRC Performance Information by Grade at Middle of Year for Study A**

| | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 |
|---|---|---|---|---|---|---|
| Sample Size (n) | | | | | | |
| Districts | 6 | 6 | 6 | 4 | 3 | 2 |
| Schools | 20 | 22 | 22 | 11 | 3 | 3 |
| Educators | 25 | 25 | 30 | 10 | 3 | 2 |
| Students | 156 | 164 | 158 | 131 | 16 | 29 |
| Gender (n) | | | | | | |
| Female | 74 | 76 | 86 | 44 | 8 | 12 |
| Male | 77 | 84 | 71 | 36 | 6 | 15 |
| Ethnicity (n) | | | | | | |
| White | 43 | 58 | 52 | 3 | 7 | 4 |
| Hispanic | 52 | 48 | 49 | 44 | 2 | 3 |
| Black | 8 | 6 | 6 | 32 | 4 | 20 |
| Native American | 0 | 0 | 0 | 1 | 0 | 0 |
| Asian | 7 | 6 | 3 | 0 | 0 | 0 |
| Multiracial | 41 | 42 | 47 | 0 | 1 | 0 |
| Other demographics (n) | | | | | | |
| Special Education | 6 | 8 | 7 | 4 | 2 | 8 |
| FRL | 19 | 38 | 22 | 26 | 5 | 18 |
| ESL | 38 | 40 | 32 | 7 | 1 | 2 |
| TRC Proficiency (n) | | | | | | |
| Far Below | 50 | 60 | 40 | 38 | 4 | 5 |
| Below | 21 | 31 | 34 | 22 | 4 | 8 |
| Proficient | 18 | 35 | 45 | 27 | 4 | 5 |
| Above | 14 | 37 | 39 | 44 | 4 | 9 |

**Experimental Design:** It is desirable that field study results are based on information from students representing the full spectrum of student ability, not just high or low performers. Therefore, participating educators were instructed to use MOY TRC results to select equal proportions of students from each performance level, based on MOY TRC results. For example, a site that contributes 16 students per grade would select four students from each performance category at MOY: Far Below Proficient, Below Proficient, Proficient, and Above Proficient.

Each student was then administered TRC using the Atlas book set with minor deviations from usual administration protocols. To ensure adequate response coverage for each text level and reduce testing burden, participating educators were instructed to administer one book at a student's MOY instructional reading level, one at the text level immediately above MOY instructional reading level, and one at the text level below. Educators were further asked to administer both Informational and Literary books to each student, and to draw from the full 76-title Atlas book set. As part of the field study, Amplify provided educators with Atlas in paper-and-pencil format. Students were also administered the appropriate measures from the Dynamic Indicators for Basic Early Literacy skills (Dynamic Measurement Group, 2010).

Following student assessment, educators entered student performance data as well as any additional comments they deemed necessary to interpret results into a secure online data collection portal. After the MOY testing period ended, educators completed a survey to provide feedback on the measures, the test administration procedures, and the study procedures. Results of this survey, along with study data, further refined the measures prior to their final implementation in software form.

**Descriptive Statistics:** Descriptive information regarding the Atlas field study results is presented in Table 2 and Table 3 by qualitative text level. Sample sizes indicating the number of students assessed by books at each text level generally decrease from PC, RB, to A through Z, which corresponds to the smaller sample sizes obtained in higher grade levels.

Student performance on the PC and RB assessments indicates that students could generally demonstrate at least two-thirds of the skills assessed. Oral reading accuracy resulting from the administration of text levels A through Z showed that students made few errors (mean = 0.95) when reading these books. Retell (for Literary books) and Recall (for Informational books) are assessed only in texts leveled A through E. Retell is scored out of 3 points — one for each of beginning, middle, and end details — with students typically providing two details (mean = 1.94). Recall is also scored out of 3 points requiring that students provide multiple key details as well as the main idea of the book. Students performed well on this measure, typically providing multiple details and the main idea (mean = 2.48). Lastly, oral comprehension is assessed in text levels D and above with five 1-point items. Average text level performance on oral comprehension ranged from 2.97 to 4.00 (overall mean = 3.45).

Table 2. Sample Size and Descriptive Statistics for PC and RB Text Levels for Study A

| Text Level | Sample Size (n) | Average | Minimum | Maximum |
|---|---|---|---|---|
| PC | 119 | 10.53 | 0 | 14 |
| RB | 138 | 4.62 | 1 | 6 |

Table 3. Sample Size and Average Accuracy, Retell, Recall, and Comprehension Performance by Text Level for Study A

| Text Level | Sample Size (n) | Accuracy | Retell | Recall | Oral Comprehension |
|---|---|---|---|---|---|
| A | 163 | 0.78 | 1.36 | 2.00 | – |
| B | 123 | 0.92 | 1.99 | 2.27 | – |
| C | 98 | 0.88 | 1.83 | 2.44 | – |
| D | 82 | 0.93 | 2.21 | 2.91 | 3.52 |
| E | 55 | 0.94 | 2.33 | 2.78 | 3.78 |
| F | 89 | 0.91 | – | – | 3.37 |
| G | 60 | 0.94 | – | – | 2.97 |
| H | 82 | 0.95 | – | – | 3.43 |
| I | 70 | 0.97 | – | – | 3.59 |
| J | 98 | 0.96 | – | – | 3.63 |
| K | 107 | 0.97 | – | – | 3.61 |
| L | 106 | 0.96 | – | – | 3.67 |
| M | 99 | 0.97 | – | – | 3.07 |
| N | 86 | 0.97 | – | – | 3.42 |
| O | 67 | 0.96 | – | – | 2.97 |
| P | 60 | 0.97 | – | – | 3.42 |
| Q | 44 | 0.97 | – | – | 3.68 |
| R | 45 | 0.98 | – | – | 3.40 |
| S | 43 | 0.98 | – | – | 3.28 |
| T | 40 | 0.97 | – | – | 3.10 |
| U | 26 | 0.97 | – | – | 3.42 |
| V | 17 | 0.98 | – | – | 3.65 |

| | | | | | |
|---|---|---|---|---|---|
| W | 15 | 0.99 | – | – | 4.00 |
| X | 27 | 0.98 | – | – | 3.11 |
| Y | 22 | 0.97 | – | – | 3.77 |
| Z | 14 | 0.98 | – | – | 3.43 |

## Study B: Print Concepts and Reading Behaviors Cut Point Development

**Purpose:** Study B was designed to determine performance expectations for the Print Concepts and Reading Behaviors tasks within TRC.

**Participants:** A subset of kindergarten students in the Atlas field study were administered Print Concepts, Reading Behaviors, and an instructional level Atlas text during the 2013–2014 middle-of- year benchmark period.

**Demographic Information:** There were 95 kindergarten students from eight schools in five districts in this analysis. The sample was composed of participants from the following demographic categories: 48 percent female, 52 percent male; 35 percent white, 49 percent Hispanic, 9 percent black, 7 percent other race; 37 percent were learning English as a second language; 44 percent were eligible for free or reduced priced lunch; 6 percent were from special education.

**Experimental Design:** To facilitate later analysis of books and items associated with Print Concepts (PC) and Reading Behaviors (RB), approximately half of participant educators administered the PC and RB books to kindergarten students instead of the one-up, one-down method implemented within the larger Atlas field study. Students were also administered the appropriate measures from the Dynamic Indicators for Basic Early Literacy Skills (Dynamic Measurement Group, 2010).

**Descriptive Statistics:** Table 4 presents descriptive information, including the sample sizes and student performance on PC and RB. The mean instructional text level for the cut points study sample is level B.

**Table 4. Sample Size and Descriptive Statistics for PC and RB Text Levels for Study B**

| Text Level | Sample Size (n) | Average | Minimum | Maximum |
|---|---|---|---|---|
| PC | 91 | 10.43 | 0 | 14 |
| RB | 94 | 4.66 | 1 | 6 |

## Study C: Inter-Rater Reliability and Alternate Form Reliability Study

**Purpose:** Study C was designed to examine the inter-rater and alternate form reliability of the Atlas book set.

**Recruitment:** Amplify educational consultants, who also acted as raters for the study, led most recruitment efforts. Schools with an interest in the study and strong existing partnerships with Amplify were asked to participate.
Participants: Three raters assessed 33 students from two schools in two Southern states during the 2013–2014 end-of-year benchmark administration period.

**Demographic Information:** Participating educators used existing data to select students by their current instructional reading level. This allowed for coverage of all text levels within Atlas. Among the students, representation was as follows: 8 from kindergarten, 10 from Grade 1, four from Grade 2, four from Grade 3, two from Grade 4, and five from Grade 5. The sample was 39 percent female and 61 percent male; 9 percent white, 21 percent Hispanic, 67 percent black, and 3 percent represented other races. The raters were two Amplify consultants.

**Experimental Design:** Educators administered four books to each participant in this study: two at the student's middle-of-year instructional reading level and two at the text level immediately above that level.

**Descriptive Statistics:** Descriptive information for the inter-rater reliability and alternate form reliability results are presented in Table 5 and Table 6 by qualitative text level. Sample sizes are provided for both the number of assessments and the number of unique students. Because students were assessed multiple times using different raters and different forms, the number of unique students does not equal the number of assessments. Similarly, the descriptive statistics for the text level are the average performance results of the assessments (including performance as determined by multiple raters on multiple forms for each student).

Student performance on the PC and RB tasks indicates that students could generally demonstrate at least two-thirds of the skills required at those text levels. Oral reading accuracy resulting from the administration of text levels A through Z showed that students made few errors (mean = 0.95) when reading these books. Retell (for literary books) and recall (for informational books) are assessed only in text levels A through E. Retell is scored out of 3 points — one each for beginning, middle, and end details — with students typically providing one detail (mean = 0.74). Recall is also scored out of 3 points and requires that students provide multiple key details as well as the main idea of the book; students typically provided one detail or the main idea (mean = 0.69). Lastly, oral comprehension is assessed in text levels D and above by five 1-point items with average text level performance ranging from 0.88 to 4.50 (overall mean = 2.24).

Table 5. Sample Size and Descriptive Statistics for PC and RB Text Levels for Study C

| Text Level | Assessments (n) | Unique Students (n) | Average | Minimum | Maximum |
|---|---|---|---|---|---|
| PC | 9 | 3 | 8.89 | 6 | 12 |
| RB | 9 | 3 | 4.33 | 2 | 6 |

Table 6. Sample Size and Average Accuracy, Retell, Recall, and Comprehension Performance by Text Level for Study C

| Text Level | Assessments (n) | Unique Students (n) | Accuracy | Retell | Recall | Comprehension |
|---|---|---|---|---|---|---|
| A | 6 | 2 | 0.83 | 0.83 | 0.00 | – |
| B | 8 | 2 | 0.94 | 0.50 | 0.20 | – |
| C | 8 | 3 | 0.83 | 0.63 | 0.00 | – |
| D | 8 | 2 | 0.95 | 0.75 | 1.50 | 3.57 |
| E | 8 | 2 | 0.93 | 1.00 | 1.75 | 2.43 |
| F | 6 | 2 | 0.93 | – | – | 1.33 |
| G | 8 | 2 | 0.96 | – | – | 2.25 |
| H | 8 | 2 | 0.92 | – | – | 1.38 |
| I | 8 | 3 | 0.90 | – | – | 2.50 |
| J | 8 | 2 | 0.96 | – | – | 2.50 |
| K | 10 | 3 | 0.91 | – | – | 1.10 |
| L | 8 | 2 | 0.87 | – | – | 0.88 |
| M | 10 | 3 | 0.97 | – | – | 1.30 |
| N | 10 | 3 | 0.97 | – | – | 2.40 |
| O | 8 | 2 | 0.98 | – | – | 0.88 |
| P | 8 | 2 | 0.97 | – | – | 2.13 |
| Q | 8 | 2 | 0.96 | – | – | 1.88 |
| R | 8 | 2 | 0.97 | – | – | 1.63 |
| S | 12 | 3 | 0.98 | – | – | 1.67 |
| T | 8 | 2 | 0.99 | – | – | 2.00 |
| U | 4 | 1 | 1.00 | – | – | 3.00 |
| V | 4 | 1 | 0.98 | – | – | 2.50 |

| | | | | | | |
|---|---|---|---|---|---|---|
| W | 4 | 1 | 0.99 | – | – | 4.50 |
| X | 4 | 1 | 0.99 | – | – | 4.00 |
| Y | 12 | 3 | 0.99 | – | – | 2.33 |
| Z | 8 | 2 | 0.99 | – | – | 3.50 |

## Study D: Additional Validity and Reliability Research for Grades 4–6

**Purpose:** Study D was designed to provide additional reliability and validity evidence for Atlas in intermediate grades (Grades 4–6) during the 2014–2015 school year for two primary reasons: 1) the preliminary field study in 2013–2014 (Study A) had a limited sample of students in Grade 4 (16 students) and Grade 5 (29 students); and 2) Amplify customers desire to use the book set with Grade 6, and study A did not include students in Grade 6. Evidence for the difficulty, internal reliability, concurrent validity, and predictive validity using DIBELS Next scores was examined in this study.

**Recruitment:** The targeted sample size was 450 students in Grades 4–6 (150 students per grade) with a minimum of five schools represented for the field study. The study used two different strategies to recruit schools: 1) Amplify Account Management and Sales teams contacted existing Amplify customers about the study directly; and 2) current Amplify customers received an informational flyer about the study via email. Two schools were enrolled in the study using the former approach, and three schools were enrolled in the study using the latter. Additionally, two schools from the same district were enrolled in the study based on an existing relationship with Belmont Abbey College, the research partner for this field study.

**Participants:** The study was conducted during the 2014–2015 middle-of-year (MOY; February– March) and end-of-year (EOY; May) benchmark administration periods. As with previous field research studies, entire classrooms of students in Grades 4–6 were recruited for participation in the study. In total, 434 students were assessed in Grade 4 (n = 133), Grade 5 (n = 140), and Grade 6 (n = 161) by 16 data collectors in seven schools, representing six districts in the states of Connecticut, North Carolina, Oklahoma, Utah, Wisconsin, and the District of Columbia. All data collectors received training in both DIBELS and TRC assessment, and demonstrated adequate inter-rater agreement prior to data collection.

**Demographic Information:** The sample was composed of participants from the following demographic categories: 46 percent female, 46 percent male, and 8 percent unspecified gender; 52 percent white, 10 percent black, 13 percent Hispanic, and 25 percent other race or race unspecified; 47 percent eligible for free or reduced lunch; 7 percent receiving special education services; and 3 percent learning English as a second language

**Experimental Design:** Each student was administered the Dynamic Indicators of Basic Early Literacy Skills (DIBELS Next; Dynamic Measurement Group, 2010) by an educator from his or her school or by a research project data collector. The DIBELS Next results were used to determine the Atlas text level that would approximate an individual student's instructional reading level. Each student was then administered TRC using the Atlas book set with minor deviations from usual administration protocols. For schools that administered DIBELS Next instead of research project data collectors, it was advised that TRC Atlas be administered no later than two weeks following DIBELS administration.[2] To ensure adequate response coverage for each text level and to minimize testing burden, data collectors administered one book at a student's instructional reading level (as determined using DIBELS Next data), one at the text level immediately above instructional reading level, and one at the text level below. Data collectors were further asked to administer both Informational and Literary books to each student, and to draw from the full 76-title Atlas book set to ensure adequate within-level book coverage. All assessments were administered using mCLASS software.

**Descriptive Statistics:** Descriptive information regarding the Study D results is presented in Table 7 and Table 8 by qualitative text level. Oral reading accuracy resulting from the administration of text levels A through Z showed that students made few errors (mean = 0.95) when reading these books, which was expected given that students were administered books at approximately their instructional reading level. Retell (for Literary books) and Recall (for Informational books) are assessed only in texts leveled A through E, which are rarely administered to students in Grades 4–6. Oral comprehension is assessed in text levels D and above with five 1-point items; in study D, average text level performance ranged from 2.95 to 5.00 (mean = 3.67).

2 Administration of DIBELS and TRC was within a two-week window for all schools in the study with the exception of one school.

**Table 7: Sample Size and Demographic Information by Grade at Middle of Year for Study D**

| | Grade 4 | Grade 5 | Grade 6 |
|---|---|---|---|
| Sample Size (n) | | | |
| Districts | 5 | 5 | 6 |
| Schools | 5 | 5 | 6 |
| Students | 133 | 140 | 161 |
| Gender (n) | | | |
| Female | 51 | 66 | 84 |
| Male | 72 | 60 | 68 |
| Ethnicity (n) | | | |
| White | 83 | 80 | 62 |
| Hispanic | 19 | 12 | 24 |
| Black | 7 | 15 | 22 |
| Native American | 9 | 13 | 9 |
| Asian | 2 | 5 | 3 |
| Multiracial | 3 | 1 | 4 |
| Other demographics (n) | | | |
| Special Education | 13 | 4 | 12 |
| FRL | 66 | 57 | 80 |
| ESL | 2 | 0 | 9 |

Table 8. Sample Size and Average Accuracy, Retell, Recall, and Comprehension Performance by Text Level for Study D

| Text Level | Sample size (n) | Accuracy | Retell/Recall | Oral Comprehension |
|---|---|---|---|---|
| C | 1 | 0.92 | 1.00 | – |
| D | 1 | 0.85 | 2.00 | 5.00 |
| E | 2 | 0.95 | 3.00 | 5.00 |
| F | 2 | 0.86 | – | 4.50 |
| G | 4 | 0.90 | – | 3.00 |
| H | 4 | 0.94 | – | 4.50 |
| I | 12 | 0.94 | – | 3.42 |
| J | 10 | 0.94 | – | 3.60 |
| K | 15 | 0.94 | – | 3.93 |
| L | 13 | 0.95 | – | 3.54 |
| M | 22 | 0.96 | – | 2.95 |
| N | 36 | 0.96 | – | 3.14 |
| O | 59 | 0.97 | – | 3.08 |
| P | 80 | 0.97 | – | 3.49 |
| Q | 109 | 0.97 | – | 3.23 |
| R | 142 | 0.97 | – | 3.50 |
| S | 132 | 0.98 | – | 3.33 |
| T | 103 | 0.97 | – | 3.16 |
| U | 72 | 0.98 | – | 3.72 |
| V | 89 | 0.98 | – | 3.92 |
| W | 106 | 0.98 | – | 3.74 |
| X | 131 | 0.99 | – | 3.37 |
| Y | 89 | 0.98 | – | 3.60 |
| Z | 62 | 0.99 | – | 3.58 |

## Study E: Additional Inter-Rater and Alternate Form Reliability Research for Grades 4–6

**Purpose:** Study E was designed to examine the inter-rater and alternate-form reliability of the Atlas book set for students in Grades 4–6 during the 2014–2015 school year for two primary reasons: 1) the preliminary inter-rater and alternate-form reliability study in 2013–2014 (Study C) had a limited sample of students in Grade 4 (2 students) and Grade 5 (5 students); and 2) Amplify customers desire to use the book set with Grade 6, which was not included in Study C.

**Recruitment:** The targeted sample size was a minimum of 10 students per grade in Grades 4–6 (30 students) with an ideal sample of 30 students per grade in Grades 4–6 from one to two schools. Recruitment was completed via an existing relationship with Belmont Abbey College, the research partner for this field study.

**Participants:** The study was conducted between the 2014–2015 beginning-of-year and middle-of- year benchmark administration periods. In total, four raters assessed 40 students from two schools in two Southern states during the 2014–2015 MOY benchmark administration period. Students in Grades 4–6 at these schools were randomly selected for participation from one classroom in each grade and signed parental consent was obtained for those students. Both schools were from the same district; one school was a K–5 elementary school and the other was a 6–8 middle school. The raters were one Belmont Abbey College professor of education and three Amplify research staff. All raters received training in both DIBELS and TRC assessment, and demonstrated adequate inter-rater agreement prior to data collection.

**Demographic Information:** The sample was composed of students in Grade 4 (n = 15), Grade 5 (n = 15), and Grade 6 (n = 10); 39 percent of the students were female and 61 percent male; 67 percent of students were black, 21 percent were Hispanic, 9 percent were white, and 3 percent were of other ethnicity.

**Experimental Design:** Each student was administered DIBELS Next by a research project data collector to determine the Atlas text level that would approximate an individual student's instructional reading level. Each student was then administered TRC using the Atlas book set with minor deviations from usual administration protocols. One data collector led the administration of the books while another shadow scored. These rater pairings and roles (direct assessor or shadow scorer) were alternated throughout the administration process. Additionally, to ensure adequate response coverage for each text level and reduce testing burden, data collectors administered two books at a student's instructional reading level (as determined using DIBELS data), and two books at the text level immediately above. Data collectors were further asked to administer both Informational and Literary books to each student, and to draw from the full 76-title Atlas book set. Within the

appropriate ranges, testers selected books for students in a manner that maximized coverage of the Atlas text levels. Table 9 provides Atlas text level ranges appropriate for students based on DIBELS Next Composite scores, as determined by the relationship of DIBELS Next and TRC from the national mCLASS database. All assessments were administered using mCLASS software.

**Table 9. Text Level Ranges Administered for Study E**

| Grade | DIBELS Composite Score Interpretation | Range of Atlas Text Levels |
|---|---|---|
| 4 | Red | H–M |
| 4 | Yellow | L–P |
| 4 | Green | O–T |
| 5 | Red | I–P |
| 5 | Yellow | P–S |
| 5 | Green | R–W |
| 6 | Red | K–R |
| 6 | Yellow | R–U |
| 6 | Green | T–Z |

**Descriptive Statistics:** Sample sizes are provided for both the number of assessments (i.e., books) administered at each level and the number of unique students who were tested at each level (Table 10). The number of assessments administered does not equal the number of unique students assessed because students were assessed with four books by two raters simultaneously, meaning that each student had eight assessment results.

Table 10. Grade and Text Levels Distributions for Study E

| Text Level | Grade 4 (n) | Grade 5 (n) | Grade 6 (n) |
|---|---|---|---|
| H | 4 | 0 | 0 |
| J | 0 | 4 | 0 |
| K | 4 | 0 | 0 |
| L | 0 | 4 | 0 |
| M | 0 | 4 | 0 |
| N | 12 | 0 | 0 |
| O | 12 | 0 | 0 |
| P | 8 | 8 | 0 |
| Q | 8 | 4 | 0 |
| R | 4 | 4 | 4 |
| S | 4 | 4 | 4 |
| T | 0 | 4 | 4 |
| U | 0 | 4 | 8 |
| V | 0 | 8 | 0 |
| W | 0 | 8 | 0 |
| X | 4 | 0 | 4 |
| Y | 0 | 0 | 12 |
| Z | 0 | 4 | 4 |

Descriptive information for the inter-rater reliability and alternate-form reliability study is presented in Table 11 by qualitative text level as the average performance results for all assessments administered at that level. For instance, one student was tested with level H books; the descriptive statistics provided for level H include the results for two level H books as scored by two raters.

Student oral reading accuracy results for text levels H through Z show that students made few errors when reading these books (accuracy mean = 0.96), which was expected given that text levels were selected for students that approximated their instructional reading level based on their DIBELS Next results. Oral comprehension performance ranged from 2.50 to 4.38 points out of a maximum of 5, and the overall mean score was 3.57. Overall book performance ranged from 0.00 to 1.75 and the

overall mean was 1.00, indicating students generally performed at their instructional text level (Note: overall book performance categories, i.e., Frustrational, Instructional, and Independent, were ordinally coded from 0 to 2 to facilitate data analysis).

**Table 11. Sample Sizes and Average Accuracy, Retell, Recall, and Comprehension Performance by Text Level for Study E**

| Text Level | Assessments (n) | Unique Students (n) | Accuracy | Comprehension | Overall Book Performance (FRU/INS/IND) |
|---|---|---|---|---|---|
| H | 4 | 1 | 0.91 | 4.25 | 0.50 |
| J | 4 | 1 | 0.97 | 2.50 | 0.00 |
| K | 4 | 1 | 0.95 | 4.25 | 1.75 |
| L | 4 | 1 | 0.96 | 4.00 | 1.50 |
| M | 4 | 1 | 0.97 | 3.75 | 1.50 |
| N | 12 | 3 | 0.98 | 3.08 | 0.67 |
| O | 12 | 3 | 0.95 | 3.67 | 0.92 |
| P | 16 | 4 | 0.97 | 3.62 | 1.00 |
| Q | 12 | 3 | 0.97 | 3.33 | 0.83 |
| R | 12 | 3 | 0.97 | 2.92 | 0.50 |
| S | 12 | 3 | 0.97 | 4.17 | 1.33 |
| T | 8 | 2 | 0.97 | 3.62 | 1.25 |
| U | 12 | 3 | 0.97 | 4.17 | 1.33 |
| V | 8 | 2 | 0.98 | 4.38 | 1.75 |
| W | 8 | 2 | 0.98 | 3.00 | 1.00 |
| X | 8 | 2 | 0.97 | 3.12 | 0.75 |
| Y | 12 | 3 | 0.96 | 3.42 | 0.83 |
| Z | 8 | 2 | 0.96 | 3.00 | 0.62 |

**Note:** Overall Book Performance FRU = 0, INS = 1, and IND = 2

# Reliability

Reliability is generally described as the consistency of a measuring instrument; reliability statistics present information about the precision of an instrument, expressed as a ratio. A test with perfect score precision has a reliability coefficient equal to 1, meaning that 100 percent of the variation among persons' scores is attributable to variation in the trait or skill the test measures, and none of the variation is attributable to error. Perfect reliability is unattainable in educational measurement; a test with a reliability coefficient of 0.90 is more likely. On such a test, 90 percent of the variation among students' scores is attributable to the trait or skill being measured, and 10 percent is attributable to errors of measurement. If the trait or skill were measured a second time, students' scores would fluctuate to some degree; that is, scores on the second test would not be perfectly consistent with the same students' initial scores.

Further, reliability is an essential characteristic of interim and formative assessments that are used for instructional decision-making; if results are spurious and unreliable, inappropriate decisions might be made. Salvia, Ysseldyke, & Bolt's (2013) standards for reliability were used to evaluate the reliability data for the Atlas book set. According to these standards, a minimum reliability of 0.60 is required to make educational decisions about groups of students, a minimum of 0.70 suggests adequate reliability generally, a minimum of 0.80 is required for screening decisions, and a minimum of 0.90 is required for important educational decisions concerning an individual student.

This section provides details on three types of reliability evidence for Atlas: internal consistency, inter- rater reliability, and alternate form reliability.

- Internal consistency reliability refers to a person's degree of confidence in the precision of scores from a single measurement.
- Inter-rater reliability estimates the degree to which different raters make consistent estimates of the same performance.
- Alternate form reliability indicates the extent to which test results generalize to different forms. Alternate forms of the test with different items should give approximately the same scores.

## Internal Consistency Reliability

Internal consistency reliability refers to one's degree of confidence in the precision of scores from a single measurement and was examined via Study A. If the test's internal consistency is 95 percent, just 5 percent of the variation of test scores is attributable to measurement error. Theoretically, this value quantifies the degree of correspondence for results across numerous administrations.
Realistically, students are not subject to multiple administrations of TRC, or any other assessment, during any single period; therefore, statistical indices of internal consistency were developed to provide evidence for the reliability of assessments.

Internal consistency reliability evidence is presented separately for PC and RB text levels and for all other (A–Z) text levels. This reflects the differences in content and structure at these text levels; PC and RB result in student scores on a number of items representing foundational reading skills

whereas text levels A–Z assess accuracy and comprehension according to oral reading performance and scoring rubrics.

### Print Concepts and Reading Behaviors

The internal consistency reliability for the Print Concepts (PC) and Reading Behaviors (RB) tasks was calculated using Cronbach's alpha for each of the three books. There are typically no missing responses in PC and RB forms so a typical indicator for internal consistency — Cronbach's alpha — is appropriate. Cronbach's alpha quantifies the degree to which the items on an assessment all measure the same underlying construct.

The results are presented in Table 12. The sample sizes are also provided for each book in the brackets. The median reliability coefficient for PC and RB is above 0.60, which meets the minimum reliability requirements appropriate for making educational decisions for groups of students according to Salvia, Ysseldyke, and Bolt (2013).

**Table 12. PC and RB Cronbach's Alpha**

| Grade | PC Book 1 | PC Book 2 | PC Book 3 | Median PC | RB Book 1 | RB Book 2 | RB Book 3 | Median RB |
|---|---|---|---|---|---|---|---|---|
| Overall | 0.70 (50) | 0.85 (32) | 0.80 (37) | 0.80 | 0.42 (45) | 0.61 (55) | 0.68 (37) | 0.61 |
| K | 0.70 (47) | 0.86 (30) | 0.80 (30) | 0.80 | 0.44 (35) | 0.62 (46) | 0.70 (31) | 0.62 |

### Atlas Text Levels A to Z

Because students are administered neither all texts nor all comprehension items in TRC, the typical indicator of internal consistency (Cronbach's alpha), is an inappropriate measure of reliability for Atlas texts levels A through Z. Therefore, marginal reliability (Sireci, Thissen, & Wainer, 1991), an appropriate reliability measure under the Item Response Theory (IRT) framework, was calculated to provide evidence for the internal consistency of the Atlas book set.

Marginal reliability requires the estimation of student ability and standard error under an IRT model, therefore, information from other editions of TRC and student performance from the national TRC database served to anchor estimation of student ability. IRT difficulty estimates were estimated for student performance beyond the field research studies and subsequently anchored to facilitate estimation of student ability and standard errors resulting from the research studies, thus allowing calculation of marginal reliability.

Students in Study A were tested with the other editions of TRC in addition to the Atlas books as part of regular classroom assessment; because IRT difficulty estimates for these other books are available in our database, estimation of student ability and standard errors resulting from administration of the Atlas books during this study was facilitated by including student performance on the non-Atlas books and anchoring on the existing difficulty values. Similarly, in Study D, the IRT difficulty estimates for the Atlas books were first obtained from operational (i.e., nonexperimental) administration of TRC during the 2014–2015 school year; these difficulty values were then anchored to allow estimation of student ability and standard errors resulting specifically from the field study.

The results of these analyses, both overall and for specific grades, are presented in Table 13 (results for Study A) and Table 14 (Study D). According to Salvia, Ysseldyke, and Bolt (2013), the minimum acceptable reliability coefficient for screening purposes is 0.80. Overall, the marginal reliability for text levels A–Z is 0.98 in both Study A and Study D. Across kindergarten to Grade 5, marginal reliability ranges from 0.90 to 0.99 (Study A); similarly, marginal reliability in Grades 4–6 is 0.98 for each grade (Study D). Therefore, evidence is provided in support of the internal consistency reliability of Atlas, indicating that text levels A–Z are a coherent set of items targeting specific reading skills.

Table 13. Marginal Reliability Results for Study A

| Grade | Unique Students (n) | Marginal Reliability |
|---|---|---|
| Overall | 653 | 0.98 |
| K | 155 | 0.93 |
| 1 | 164 | 0.97 |
| 2 | 158 | 0.93 |
| 3 | 131 | 0.96 |
| 4 | 16 | 0.99 |
| 5 | 29 | 0.90 |

Table 14: Marginal Reliability Results for Study D

| Grade | Unique Students (n) | Marginal Reliability |
|---|---|---|
| Overall | 434 | 0.98 |
| 4 | 133 | 0.98 |
| 5 | 140 | 0.98 |
| 6 | 161 | 0.98 |

## Inter-Rater Reliability

In observational assessments such as TRC, it is important that student performance be unrelated to or unaffected by a specific test administrator. Because there is a degree of subjectivity in scoring the accuracy of oral reading and comprehension, it is important to examine the degree to which TRC administrators can score student reading accuracy in a standardized and consistent manner. The sources of error associated with inter-rater reliability lie in the assessor.

In presenting inter-rater reliability (IRR) evidence, raters' scores are typically compared using either Cohen's kappa or intraclass correlations (ICC). Kappa indicates the degree of agreement between raters on nominal or categorical variables; ICC is one of the most commonly used statistics for assessing IRR on ordinal, interval, or ratio variables and is suitable for studies with two or more coders (Hallgren, 2012). Fleiss (1981) suggested kappa values greater than 0.75 to indicated excellent agreement, 0.40 to 0.75 as fair to good, and below 0.40 as poor. Cicchetti (1994) provides cutoffs for ICC values, with IRR being poor for values less than 0.40, fair for values between 0.40 and 0.59, good for values between 0.60 and 0.74, and excellent for values between 0.75 and 1.00.

IRR estimates reported here are based on two or more independent assessors simultaneously scoring student performance during a single test administration ("shadow-scoring"). Reliability coefficients presented therefore represent the degree to which the administration and scoring procedures for the TRC components lead to consistent results, generalizing across administrators.

The IRR results for overall book performance, reading record accuracy, oral comprehension performance, and recall/retell performance in kindergarten through Grade 5 as resulting from Study C are presented overall and by grade in Table 15. Grade-specific IRR is slightly lower than overall IRR due to relatively small sample sizes in each of the grades. All values are above 0.40, with the exceptions of overall book performance in Grade 2 and reading record accuracy in Grade 3. According to Cicchetti's criteria, IRR overall is classified as good for overall book performance and excellent for reading record accuracy, oral comprehension, and retell/recall.

**Table 15. Inter-Rater Reliability Results for Study C**

| Grade | Book Performance (FRU/INS/IND) | Reading Record Accuracy | Oral Comprehension | Retell/Recall |
|---|---|---|---|---|
| Overall | 0.67 | 0.94 | 0.74 | 0.80 |
| K | 0.77 | 0.97 | 0.47 | 0.85 |
| 1 | 0.49 | 0.98 | 0.68 | 0.64 |
| 2 | 0.35 | 0.57 | 0.72 | N/A |
| 3 | 0.99 | 0.16 | 0.58 | N/A |
| 4 | 0.99 | 0.82 | 0.99 | N/A |
| 5 | 0.74 | 0.53 | 0.68 | N/A |

**Note:** Retell/recall is not typically administered in Grade 2 or above.

The IRR results for Grades 4–6 resulting from Study E are presented in Table 16, including overall book performance, reading record accuracy, oral comprehension performance, and recall/retell performance for the entire sample and for each grade. All values are above 0.40, except for overall book performance for Grade 4 and reading record for Grade 6. According to Cicchetti's criteria, IRR overall is classified as fair for overall book performance, excellent for reading record accuracy, and good for oral comprehension. Cohen's kappa was also explored and found to suggest fair-to-good inter-rater reliability overall and for grade-specific performance. Further analysis of rater results found that raters achieved perfect agreement on 78 percent of administrations for overall book performance classifications (i.e., FRU/

INS/IND), 95 percent of reading record accuracy classifications, and 78 percent of oral comprehension classifications. These classifications are the TRC results that are most relevant to instructional decision-making, and the strong inter-rater agreement found on student classifications lends confidence in the reliability of using standardized TRC assessment procedures to characterize student reading achievement and use that information to guide instruction.

**Table 16. Inter-Rater Reliability Results for Study E**

| Grade | Book Performance (ICC) | Book Performance (Kappa) | Reading Record Accuracy (ICC) | Oral Comprehension (ICC) |
|---|---|---|---|---|
| Overall | 0.42 | 0.59 | 0.89 | 0.61 |
| 4 | 0.37 | 0.48 | 0.94 | 0.67 |
| 5 | 0.48 | 0.63 | 0.58 | 0.63 |
| 6 | 0.31 | 0.59 | 0.91 | 0.44 |

## Alternate Form Reliability

Alternate form reliability indicates the extent to which test results generalize to different forms of the assessment content. To demonstrate alternate form reliability, students were tested with two different (i.e., alternate) but equivalent forms of the test, and scores from these two forms were correlated. Student learning and potential practice could lead to differing scores on parallel test forms. However, alternate forms of the test with different items should yield approximately the same scores.

### Atlas Text Levels A to Z

There are two to three books per text level in Atlas; books at the same text level are considered alternate forms. An individual student's performance on the alternate books at the same text level should yield approximately the same scores on oral reading accuracy, comprehension, and/or retell/recall, as well as overall book performance.

Study C provided alternate form reliability data for administrations in kindergarten through Grade 5; 27 students were each assessed on two texts at their instructional reading level and two texts of the one level below their instructional reading level. Each component of TRC (accuracy, comprehension, retell/recall) as well as overall book performance was submitted to paired t-test comparisons. Table 17 presents the detailed t-test results.

Table 17. Alternate Form Reliability Results from Study C

| Grade | Accuracy | Retell/Recall | Oral Comprehension | Overall Book Performance |
|---|---|---|---|---|
| All | t(45) = –0.30, n.s. | t(7) = 0.88, n.s. | t(41) = –0.58, n.s. | t(47) = –0.74, n.s. |
| K | t(5) = 0.04, n.s. | t(4) = 0.23, n.s. | t(2) = 1.51, n.s. | t(7) = 1.93, n.s. |
| 1 | t(12) = –0.08, n.s. | t(2) = 1.73, n.s. | t(11) = –0.20,n.s. | t(12) = –0.56, n.s. |
| 2 | t(5) = 1.87, n.s. | N/A | t(5) = 1.75, n.s. | t(5) = 1.00, n.s. |
| 3 | t(6) = –1.54, n.s. | N/A | t(6) = –0.68, n.s. | t(6) = –1.00, n.s. |
| 4 | t(3) = 0.29, n.s. | N/A | t(3) = –2.32, n.s. | t(3) = –1.00, n.s. |
| 5 | t(9) = 0.00, n.s. | N/A | t(9) = –2.38, $p < 0.05$ | t(9) = –2.45, $p < 0.05$ |

Study E provided alternate-form reliability evidence for 40 students in Grades 4–6. Each component of TRC (accuracy and comprehension only; retell/recall is not available for books at levels typically appropriate for students in Grades 4–6) as well as overall book performance was submitted to paired t-test comparisons and the results are presented in Table 18.

Table 18. Alternate Form Reliability Results from Study E

| Grade | Accuracy | Retell/Recall | Oral Comprehension | Overall Book Performance |
|---|---|---|---|---|
| All | t(79) = –0.03, n.s. | N/A | t(79) = 0.49, n.s. | t(79) = 1.88, n.s. |
| 4 | t(29) = 1.20, n.s. | N/A | t(29) = –0.82, n.s. | t(29) = 0.30, n.s. |
| 5 | t(29) = 1.43, n.s. | N/A | t(29) = 2.47, $p < 0.05$ | t(29) = 3.64, $p < 0.01$ |
| 6 | t(19) = –2.08, n.s. | N/A | t(19) = –0.21, n.s. | t(19) = 0.06, n.s. |

Across the entire sample, differences in overall book performance, reading record accuracy, comprehension, and retell/recall scores among alternate books within the same level administered to the same participants were nonsignificant. This suggests good alternate form reliability for all the TRC components in the Atlas book set. Examining differences by TRC component and grade finds few significant results. Specifically, significant differences are found in Study C at Grade 5 for oral comprehension, t(9) = –2.38, $p < 0.05$, and overall book performance, t(9) = –2.45, $p < 0.05$; similarly, Grade 5 results in Study E indicate differences in oral comprehension, t(29) = 2.47, $p < 0.05$, and overall book performance, t(29) = 3.64, $p < 0.01$.[2]

2 Lower alternate-form reliability in Grade 5 can be attributed to small sample sizes, and therefore increased measurement error, and to differences in the structure and language used in one pair of books; this issue will be explored in further depth to determine effective use and scoring of these books.

### Print Concepts and Reading Behaviors

To examine alternate form reliability for Print Concepts and Reading Behaviors, multiple Analyses of Variance (ANOVA) analyses were conducted using the results from Study B to test whether the performance on the total score was significantly different across the three alternative books. The results are nonsignificant for PC, $F(1,118) = 0.93$, n.s., and RB, $F(1,134) = 2.65$, n.s., indicating the total score performance was not different when administering different books for these tasks.

# Atlas Book Set Difficulty

## Text Difficulty

The overall difficulty of each text in the Atlas book set (i.e., Book Performance) as well as the difficulty of each component (i.e., reading record accuracy, recall/retell, oral comprehension) in each book was examined via an Item Response Theory (IRT; Embretson & Reise, 2000) framework. IRT attempts to quantitatively model the results of a student with a specific level of ability answering a specific question. Calibration of an IRT model results in parameter estimates of difficulty for each test item as well as estimates of student ability, placing both on a common scale that enables direct comparisons. For the purpose of examining the difficulty of texts in the Atlas book set, books were analyzed as items within an IRT framework.

The sample employed in this analysis includes all 275,527 students who were assessed using the Atlas book set during the 2014–2015 school year as part of operational administration of the assessment. Demographic information for this sample is provided in Table 19. When specified or indicated, the majority of students were enrolled in kindergarten through Grade 2 (28.86%, 30.12%, and 28.73% respectively) with smaller percentages in Grades 3–6; 43 percent were identified as female and 45 percent as male; 26 percent identified as white; 53 percent identified as native English speaking; 56 percent identified as not currently receiving special education services; and 38 percent identified as eligible for free/reduced lunch.

**Table 19. Sample Size and Demographics by Grade for TRC Atlas 2014–2015**

| Student Demographic | Level | Overall | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Grade 6 |
|---|---|---|---|---|---|---|---|---|---|
| Sample Size (n) | States | 22 | 21 | 21 | 20 | 17 | 14 | 13 | 9 |
| | Districts | 252 | 247 | 247 | 244 | 100 | 55 | 44 | 14 |
| | Schools | 1,452 | 1,216 | 1,274 | 1,243 | 505 | 325 | 262 | 28 |
| | Educators | 12,966 | 3,614 | 3,952 | 3,768 | 1,245 | 717 | 603 | 51 |
| | Students | 275,527 | 77,768 | 81,133 | 77,585 | 18,545 | 10,731 | 9,177 | 588 |
| Gender (%) | Female | 42.63% | 40.33% | 42.60% | 43.51% | 44.87% | 45.16% | 46.87% | 49.49% |
| | Male | 44.74% | 42.58% | 44.34% | 45.81% | 48.22% | 47.04% | 47.73% | 44.39% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ethnicity (%) | White | 25.63% | 24.83% | 26.69% | 28.45% | 22.09% | 17.56% | 17.10% | 6.12% |
| | Hispanic | 21.21% | 17.09% | 18.41% | 20.55% | 33.49% | 38.65% | 38.14% | 68.88% |
| | African American | 21.88% | 19.91% | 21.71% | 21.46% | 26.66% | 27.07% | 28.90% | 3.57% |
| | Native American | 0.40% | 0.51% | 0.40% | 0.36% | 0.31% | 0.25% | 0.22% | 0.34% |
| | Asian or Pacific Islander | 2.29% | 2.00% | 2.09% | 2.04% | 2.87% | 4.28% | 5.23% | 1.19% |
| | Multiracial | 2.67% | 2.57% | 2.61% | 3.06% | 3.25% | 1.31% | 1.41% | 0.85% |
| | Not Specified | 25.92% | 33.09% | 28.09% | 24.08% | 11.33% | 10.88% | 9.01% | 19.05% |
| Other Demographics (%) | Special Education | 6.96% | 4.49% | 5.86% | 7.30% | 12.33% | 13.34% | 16.14% | 13.10% |
| | Free/ Reduced Lunch | 37.35% | 33.13% | 37.10% | 39.49% | 41.61% | 42.03% | 45.29% | 4.42% |
| | English as a Second Language | 6.34% | 4.50% | 5.74% | 5.72% | 11.25% | 14.66% | 11.94% | 19.56% |
| English Language Learner | English as a Second Language | 6.34% | 4.50% | 5.74% | 5.72% | 11.25% | 14.66% | 11.94% | 19.56% |
| | Native English | 52.67% | 48.56% | 49.94% | 53.31% | 63.55% | 64.93% | 69.46% | 60.20% |
| | Not Specified | 40.99% | 46.94% | 44.32% | 40.98% | 25.20% | 20.41% | 18.60% | 20.24% |
| Special Education | Yes | 6.96% | 4.49% | 5.86% | 7.30% | 12.33% | 13.34% | 16.14% | 13.10% |
| | No | 55.61% | 50.36% | 53.64% | 54.14% | 71.40% | 75.02% | 74.62% | 67.35% |
| | Not Specified | 37.43% | 45.15% | 40.49% | 38.57% | 16.27% | 11.65% | 9.24% | 19.56% |
| Free or Reduced Lunch Status | Eligible | 37.35% | 33.13% | 37.10% | 39.49% | 41.61% | 42.03% | 45.29% | 4.42% |
| | Not Eligible | 16.96% | 15.32% | 18.36% | 18.50% | 15.00% | 13.91% | 13.78% | 6.46% |
| | Not Specified | 45.68% | 51.55% | 44.54% | 42.01% | 43.39% | 44.06% | 40.93% | 89.12% |

Performance on all texts in the Atlas book set administered to students in the aforementioned sample was considered in these analyses, not just the text associated with final or instructional reading level performance. Therefore, each student at each benchmark period could have several responses. IRT analysis was conducted across these student responses for each component of the overall assessment: book performance, reading record accuracy, retell/recall, and oral comprehension.

Although IRT encompasses a family of statistical models, the Partial Credit Rasch model (Masters, 1982) was selected both for its simplicity and its ability to accurately model student performance on TRC since each of the components as well as overall book performance are scored with successive integers. For example, overall book performance of "Frustrational" is coded as 0, "Instructional" is 1, and "Independent" is 2. The analyses were conducted in Winsteps Rasch calibration software (Linacre, 2014). Estimated difficulty values for each text and TRC component are presented in Appendix 2.

Figures 1–4 show the IRT estimated difficulties by text level for overall book performance and each component of TRC for the Atlas book set. The Atlas book set demonstrates a clear progression of difficulty corresponding to text level with limited variability of book difficulty within each text level.

A one-way Analysis of Variance (ANOVA) was conducted in order to test whether difficulty is significantly different across text levels, based on overall book performance as well as each of the components of TRC individually. Significant effects of text level were found overall and for all TRC components: overall book performance, $F(25,47) = 4661.14$, $p < 0.01$, oral reading accuracy, $F(25,47) = 672.98$, $p < 0.01$, retell/recall, $F(4,10) = 211.06$, $p < 0.01$, and oral comprehension, $F(22,41) = 6271.87$, $p < 0.01$. These results support the idea that IRT difficulty differs across text levels. Visual inspection of Figures 1–4 suggests that these significant differences correspond to a general increase in difficulty across text levels, a central characteristic of leveled book sets.

Figure 5 provides a summary of mean text level difficulties for each of the TRC components for the Atlas book set. There is a clear trend that as text level increases, the difficulties of all the components also increase.

Figure 1. Overall Book Performance Difficulty

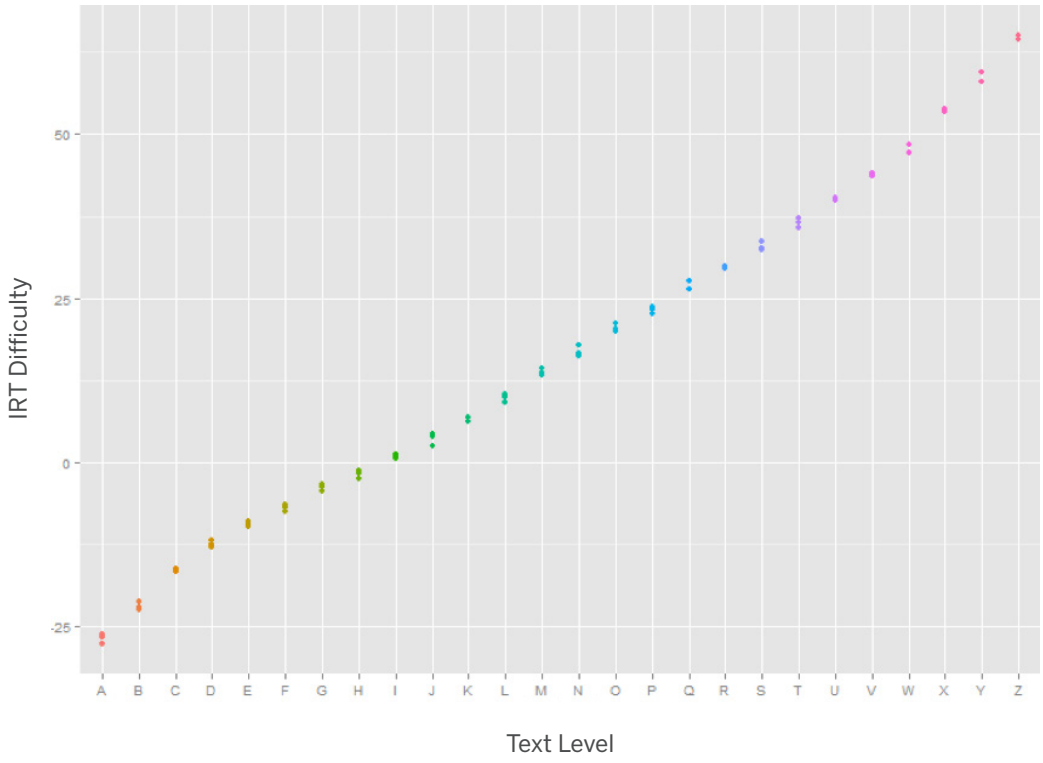



Figure 2. Reading Record Accuracy Difficulty

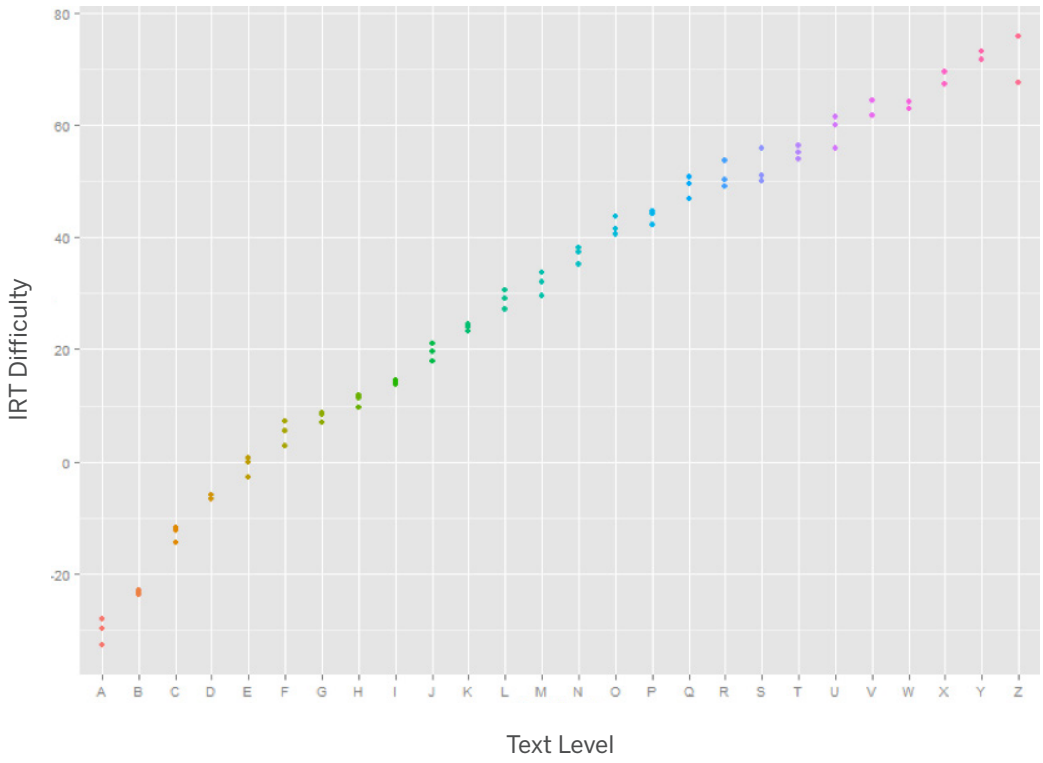


**Figure 3. Retell/Recall Difficulty**

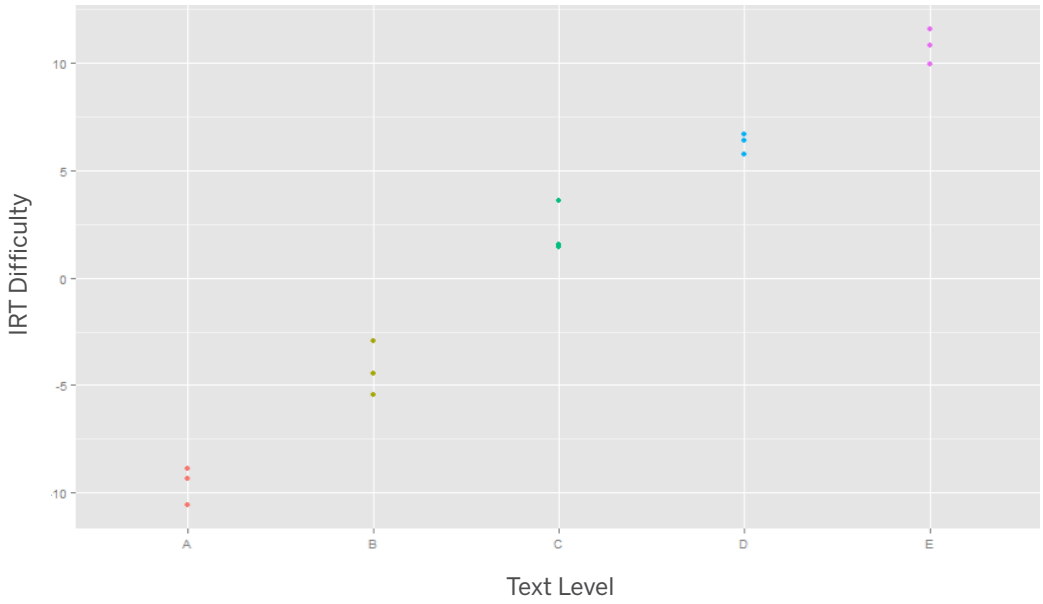



**Figure 4. Oral Comprehension Difficulty**

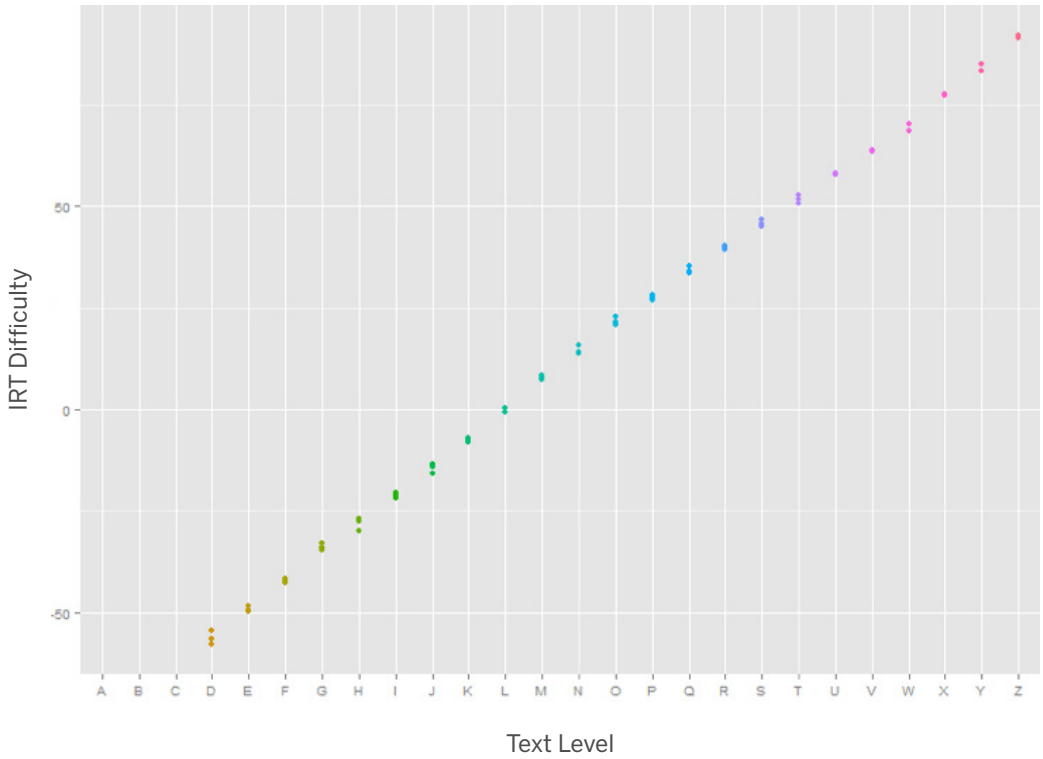
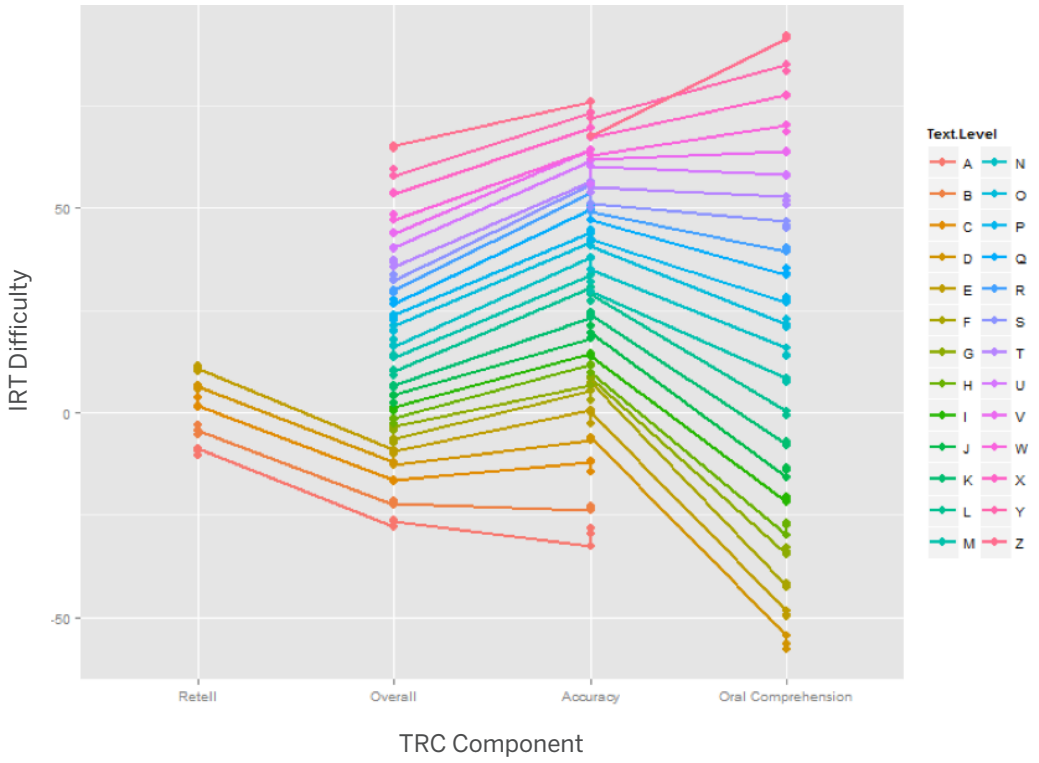


Figure 5. Summary of Difficulties by Text Level




## Book Equivalence

There are two to three books available at each text level in the Atlas book set, including a combination of literary and informational books. When determining reading record accuracy or oral comprehension using Atlas, it should not matter which book (within a given text level) is administered to a student. Stated differently, student performance should, on average, be equivalent, irrespective of which book they encounter within a given text level. Book equivalency analysis examines the consistency or reliability of Atlas results over different books within a given text level.

Analysis of book equivalence within a given text level was conducted according to the IRT procedures described previously. The specific results — including minimum, maximum, and mean difficulty for overall book performance as well as each TRC component of the Atlas book set — are presented in Appendix 3. The results suggest a very narrow range of book difficulties within text levels, and that the difficulties increase as text levels increase.

Intraclass correlation (ICC) refers to a set of coefficients representing the relationship between variables of the same class and can describe how strongly units in the same

group resemble each other. Variables of the same class share a common metric and variance, which generally means that they measure the same thing (Shrout & Fleiss, 1979). ICC analyses, therefore, describe the degree of similarity between Atlas books at similar text levels. Results of ICC analysis for student performance on the Atlas book set indicate that books within the same text level are of near-equivalent difficulty with respect to overall book performance (ICC = 0.99) as well as for reading record accuracy (ICC = 0.99), retell/recall (ICC = 0.99), and oral comprehension (ICC = 0.99). This suggests that the books at each text level are equivalent and interchangeable, which is consistent with evidence presented in the alternate form study.

Additionally, a one-way ANOVA was conducted for each of the components of the Atlas book set, examining the IRT difficulty by genre (i.e., literary or informational). The results suggest no significant effect of genre on difficulty with respect to overall book performance, $F(1,46) = 2.89$, n.s., reading record accuracy, $F(1,46) = 0.91$, n.s., and oral comprehension, $F(1,40) = 2.89$, n.s. A significant effect of genre was discovered for retell/recall performance, $F(1,9) = 19.66$, $p < 0.01$, indicating that informational books are more difficult to retell/recall than literary books. This result is likely due to both differences in text structure (literary books tend to have a more narrative structure which facilitates memory over informational text structures) and to exposure: younger students are more commonly exposed to narrative than expository texts and their structures (Duke, 2000; Yopp & Yopp, 2000; Hoffman, et al., 1994), and experience may be an important factor in genre-specific comprehension (Kamberelis, 1998; Kamil & Lane, 1997).

## Print Concepts and Reading Behaviors Item Difficulty and Fit Statistics

Student responses to items associated with Print Concepts (PC) and Reading Behaviors (RB) texts resulting from Study A were analyzed according to an IRT framework. In addition to difficulty estimates for each item, IRT provides standard error information as well as statistics indicating the degree to which an item fits the theoretical expectations of the model. Infit statistics are sensitive to students' unexpected patterns of observations on items roughly targeted at their ability (and vice- versa) while outfit statistics are sensitive to students' unexpected observations on items that are relatively easy or hard for them (and vice-versa). Expected values for both infit and outfit statistics are between 0.5 and 1.5 (Linacre, 2014); values outside this range suggest the item might distort or degrade the measurement system.

In addition, adjusted point-biserial correlations are provided for each item. The point-biserial correlation (or the point-polyserial correlation) is the Pearson correlation between item-level performance and test-level performance or raw score. The

adjusted point-biserial correlation is simply the point-biserial correlation excluding the current observation from the raw score. These are crucial for evaluating whether the coding scheme and student responses accord with the requirement that "higher observations correspond to more of the latent variable" (and vice-versa; Linacre, 2014). The higher the value, the more consistent the item is with the test. Correlations lower than 0.2 suggest that an item is not consistent, and indicates that students with high ability score low on this item.

Item difficulty, standard error, infit, outfit, and point-biserial results for PC and RB are presented in Appendix 3 and suggest that all the items in the PC and RB tasks perform well. In addition, ANOVAs were conducted to test whether the items' difficulties were significantly different across alternative books in Print Concepts and Reading Behaviors. The results are nonsignificant for both PC, $F(1,40) = 0.27$, n.s., and RB, $F(1,16) = 0.001$, n.s. — indicating that item difficulties are equivalent across books within PC and RB tasks.

# Print Concepts and Reading Behaviors Cut Points

Cut points for the Print Concepts (PC) and Reading Behaviors (RB) tasks in the Atlas edition of TRC were developed using the contrasting groups and borderline standard-setting methodologies (Cizek & Bunch, 2007) with composite score interpretations resulting from concurrent administration of DIBELS: Next (Dynamic Measurement Group, 2010). The contrasting groups method identifies students who are proficient or nonproficient readers on PC and RB according to DIBELS Next composite score interpretation (i.e., green as proficient, red as nonproficient; yellow is excluded as a borderline group). The contrasting groups method assumes that students performing Above Benchmark (green) on the DIBELS Next composite also have a greater probability of being Proficient on PC or RB while students Well Below Benchmark (red) on the DIBELS Next composite have a lower probability of being Proficient on PC or RB. Students Below Benchmark (yellow) on the DIBELS Next composite are in the borderline group.

The DIBELS measures included in calculation of the Composite score and PC and RB tasks assess basic early literacy skills that are necessary for accurate and fluent reading of connected text. Thus, the approach described above led to the development of PC and RB cut points that help educators determine whether students possess key knowledge about text and reading that is necessary for successful reading at text level A and above. Students who do not meet the cut point on either PC or RB would not be expected to read connected text accurately and fluently.

Study B provided data for the PC and RB cut points analysis. There were 95 kindergarten students from eight schools in five districts included in this analysis. Among them, 91 were tested on PC, and 94 were tested on RB. The contrasting groups method identified 57 students as Proficient and 19 as Non-Proficient based on DIBELS Next composite score interpretation, and excluded 19 students in the borderline group. Score distributions for each group are plotted on the same graph and the intersection point of the two distribution curves suggests cut points. The borderline method identified students with scores in the yellow in the DIBELS Next performance interpretation (i.e., "at the borderline") and used their mean and median performance as a reference to set cut points. Results are presented separately for PC and RB.

Figure 6 shows the distribution of student performance for the two contrasting groups on PC. The median PC score for the borderline group is 11, and the mean is 10.84. Figure 7 shows the distribution of student performance on the two contrasting groups on RB. The median RB score for the borderline group is 5, and the mean is 4.26. The results of the contrasting group and borderline methods both suggest PC scores of 11 or above to be proficient and RB scores of 5 or 6 to be proficient.

**Figure 6. Print Concepts Performance of Contrasting Groups**

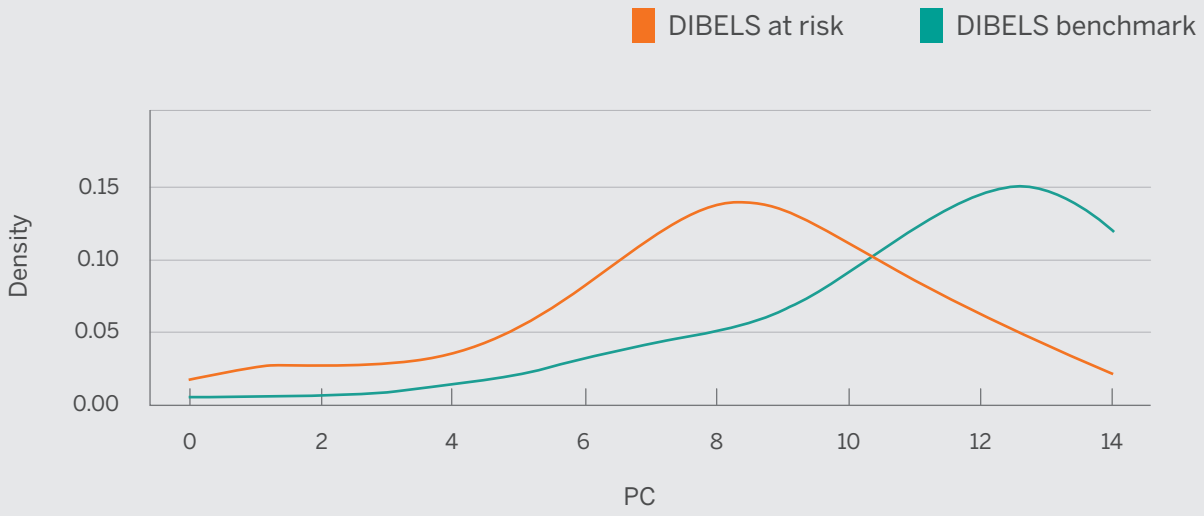
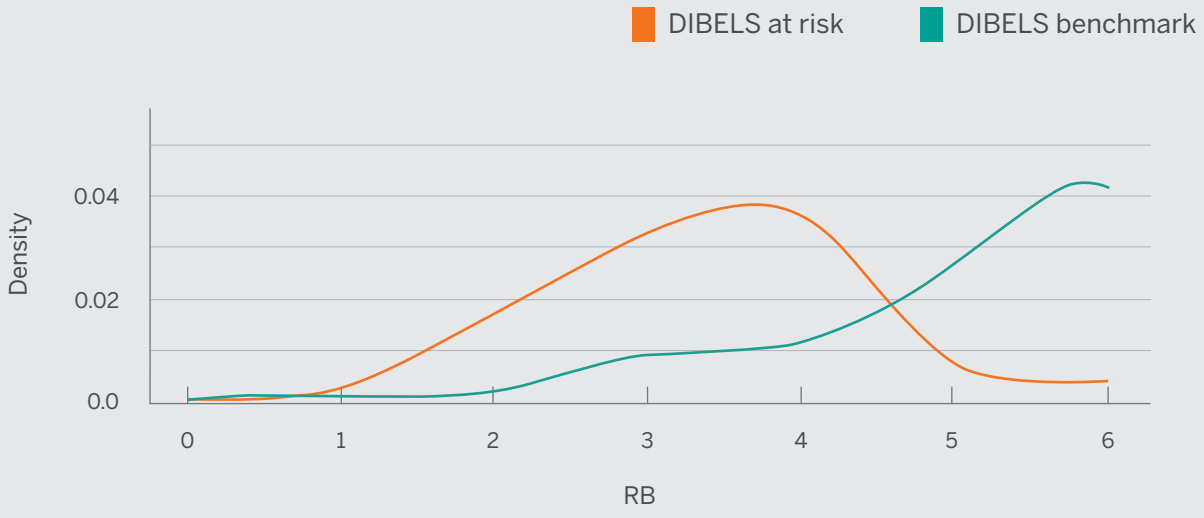



**Figure 7. Reading Behaviors Performance of Contrasting Groups**



Additionally, logistic regression and Receiver Operating Characteristic (ROC) analyses were conducted to compare the accuracy, sensitivity, specificity, and logistic prediction results for each possible cut point for PC and RB. Tables 20 and 21 show the results for PC and RB respectively. Potential cut points were identified based on considerations of accuracy, sensitivity, and specificity together. According to these analyses, 10 or 11 is an appropriate Print Concepts cut point, and 5 is an appropriate RB cut point.

Since logistic regression methods of setting cut points tend to depress cut points due to measurement error (Cizek & Bunch, 2007) and to maintain consistency with the contrasting groups results, the final PC cut point was set at 11 and the final RB cut point at 5. Therefore, students who score 11 or above on Print Concepts have an 80 percent or greater probability of being at or above benchmark on the DIBELS Composite Score; and students who scored 5 or above on Reading Behaviors have an 87 percent or greater probability of being at or above benchmark on the DIBELS Composite Score.

**Table 20. Print Concepts Total Score and Results of Logistic Regression and ROC Analysis**

| Total Score | ROC Accuracy | ROC Sensitivity | ROC Specificity | Logistic Regression Likelihood of Success | Percent of Students At/ Above Score |
|---|---|---|---|---|---|
| 0 | 0.74 | 1.00 | 0.00 | 0.12 | 1.39 |
| 1 | 0.72 | 0.98 | 0.00 | 0.16 | 1.39 |
| 2 | 0.74 | 0.98 | 0.05 | 0.21 | 1.39 |
| 4 | 0.75 | 0.98 | 0.11 | 0.32 | 1.39 |
| 5 | 0.74 | 0.96 | 0.11 | 0.39 | 2.78 |
| 6 | 0.74 | 0.94 | 0.16 | 0.47 | 2.78 |
| 7 | 0.76 | 0.94 | 0.26 | 0.54 | 6.94 |
| 8 | 0.69 | 0.85 | 0.26 | 0.62 | 9.72 |
| 9 | 0.76 | 0.83 | 0.58 | 0.69 | 8.33 |
| 10 | 0.76 | 0.77 | 0.74 | 0.75 | 4.17 |
| 11 | 0.75 | 0.74 | 0.79 | 0.80 | 15.28 |
| 12 | 0.65 | 0.57 | 0.89 | 0.85 | 12.50 |
| 13 | 0.56 | 0.42 | 0.95 | 0.88 | 16.67 |
| 14 | 0.42 | 0.21 | 1.00 | 0.91 | 15.28 |

Table 21. Reading Behaviors Total Score and Results of Logistic Regression and ROC Analysis

| Total Score | ROC Accuracy | ROC Sensitivity | ROC Specificity | Logistic Regression Likelihood of Success | Percent of Students At/ Above Score |
|---|---|---|---|---|---|
| 1 | 0.76 | 1.00 | 0.00 | 0.07 | 1.33 |
| 2 | 0.75 | 0.98 | 0.00 | 0.18 | 4.00 |
| 3 | 0.79 | 0.98 | 0.17 | 0.41 | 16.00 |
| 4 | 0.79 | 0.88 | 0.50 | 0.68 | 17.33 |
| 5 | 0.83 | 0.79 | 0.94 | 0.87 | 18.67 |
| 6 | 0.64 | 0.54 | 0.94 | 0.95 | 42.67 |

## Impact of PC and RB Cut Points

To further explore the impact of setting the PC cut point to 11 and the RB cut point to 5, student performance on text level A was analyzed. Using the newly established PC and RB cut points, students were categorized as proficient or nonproficient and mean performance on each of the TRC components at text level A was calculated. Table 22 summarizes detailed mean performance for each component at text level A. The mean performance for proficient students was higher than the nonproficient students' mean performance on both the PC and RB tasks, suggesting the cut points appropriately capture the increasing reading demands the PC and RB tasks represent.

Table 22. Mean Performance on Level A Components by Proficiency on PC and RB

| | RR Accuracy on Level A | Retell/Recall on Level A | Book Performance on Level A |
|---|---|---|---|
| PC Proficient | 78.61% | 1.42 | 0.18 |
| PC Non-Proficient | 51.48% | 0.59 | 0.06 |
| RB Proficient | 83.18% | 1.54 | 0.32 |
| RB Non-Proficient | 53.00% | 0.64 | 0.03 |

**Note:** RR Acuracy range = 0–100%); Retell/Recall range = 0–3; FRU/INS/IND coded 0/1/2.

# Validity

The validity of a test is the degree to which it assesses what it claims to measure. Formally, validity is defined as the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (American Educational Research Association, Psychological Association, & National Council on Measurement in Education, 1999). In other words, validity represents how confident we are that interpretations of test scores accurately represent what we believe they do (e.g., high scores on a comprehension assessment actually represent high comprehension skill). In this sense, validity is a way to describe a test's accuracy or utility.

Validity is not "proven" but rather evidence is collected to strengthen the assertion that a test accurately measures the desired construct(s). Validity was traditionally considered a property assessments themselves possessed; it was categorized as content-, construct-, and criterion validity. The current view, however, considers a more unified treatment under which validity evidence is collected to support test score interpretations for their intended or unintended use (Kane, 2001; Messick, 1989) and may be captured under a more general heading of evidence for construct validity. Determining the validity of a test involves the use of data and other information, both internal and external to the test instrument itself.

To facilitate discussion and demonstration, evidence for the construct validity and criterion validity of the Atlas edition of TRC is presented via concurrent and prediction results.

- Criterion-related validity is the extent to which student performance on the assessment procedure being validated can estimate student performance on a criterion measure (Salvia, Ysseldyke, & Bolt, 2013). Criterion-related validity includes concurrent and predictive validity. Evidence for the concurrent or predictive validity of an assessment refers to the degree to which current outcomes are associated with outcomes on an external, conceptually related, instrument administered near-concurrently (concurrent validity evidence) or subsequently (predictive validity evidence).
- Construct validity investigates the extent to which a test measures the construct that it claims to assess.

## Concurrent Validity

Evidence of concurrent validity is often presented as a correlation between the assessment and an external criterion measure. Instructional reading levels determined from the administration of the Atlas edition of TRC should correlate highly with other accepted procedures and measures that determine overall reading achievement, including accuracy and comprehension. The degree of correlation between two conceptually related, concurrently administered tests suggests the tests measure the same underlying psychological constructs or processes.

DIBELS Next (Dynamic Indicators of Basic Early Literacy Skills; Dynamic Measurement Group, 2010) is a set of measures used to assess early literacy and reading skills, including phonemic awareness, basic phonics, accurate and fluent reading of connected text, and reading comprehension for students from kindergarten through Grade 6. It includes the following measures: Letter Naming

Fluency (LNF), First Sound Fluency (FSF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), Oral Reading Fluency (DORF), and Daze — a maze task. An overall composite score is calculated based on a student's scores on grade-specific measures to provide an overall indication of literacy skills. DIBELS Next is considered an appropriate criterion measure given the strong reliability and validity evidence demonstrated by various studies (please refer to the DIBELS Next technical manual for details: Good et al., 2013). The relationship between final instructional reading level achieved on the Atlas edition of TRC and Composite score resulting from administration of DIBELS Next within a two-month period provides concurrent validity evidence for the Atlas
book set.

Of the 655 students who participated in Study A, final instructional reading levels were determined for 281 students, employing the same accuracy and comprehension criteria applied in nonexperimental administrations. Appendix 4 describes the specifics of these calculations. DIBELS Next Composite scores were available for 269 of those 281 students. This sample was composed of students in the following demographic categories: 48 percent female, 45 percent male, 7 percent unknown gender; 10 percent black, 33 percent white, 36 percent Hispanic, 21 percent other race or unknown race.
Table 23 summarizes the concurrent validity evidence for the entire sample and for each grade. Across grades, final instructional reading level on Atlas demonstrated moderate to strong correlations with DIBELS Next composite score, ranging 0.50 to 0.82 with an overall correlation of 0.81.
Correlation with the DIBELS Next composite score is slightly lower in kindergarten, possibly because text levels at the lower grades are much less variable due to the floor effect at kindergarten.

Correlations at Grades 4 and 5 are also slightly lower, which may be attributed to the small sample sizes at those grades. When sample size is adequate (i.e., Grades 1–3), the correlations are greater than 0.7, suggesting adequate evidence for the concurrent validity of the Atlas version of TRC with DIBELS Next.

**Table 23. Concurrent Validity Evidence for Study A**

| Grade | Students (n) | Correlation With DIBELS Next Composite Score |
|---|---|---|
| All | 269 | 0.81 |
| K | 111 | 0.63 |
| 1 | 55 | 0.78 |
| 2 | 52 | 0.82 |
| 3 | 38 | 0.77 |
| 4 | 4 | 0.50 |
| 5 | 8 | 0.55 |

Final instructional reading levels were also determined for 154 students of the total 434 students who participated in Study D, employing the criteria described previously (Appendix 4). Further, DIBELS Next Composite scores were available for 137 of those 154 students. This final sample was composed of students in the following demographic categories: 49 percent female, 44 percent male, 7 percent unknown gender; 9 percent black, 53 percent white, 12 percent Hispanic, and 23 percent other or unknown ethnicity.

Table 24 summarizes the concurrent validity evidence for the entire sample and for each grade within Study D. Across grades, final instructional reading level on Atlas demonstrated strong correlations with DIBELS Next composite score, ranging 0.95 to 0.97 with an overall correlation of 0.93. The correlations are greater than 0.7, suggesting adequate evidence for the concurrent validity of the Atlas version of TRC with DIBELS Next.

**Table 24. Concurrent Validity Evidence for the Atlas 4–6 Study**

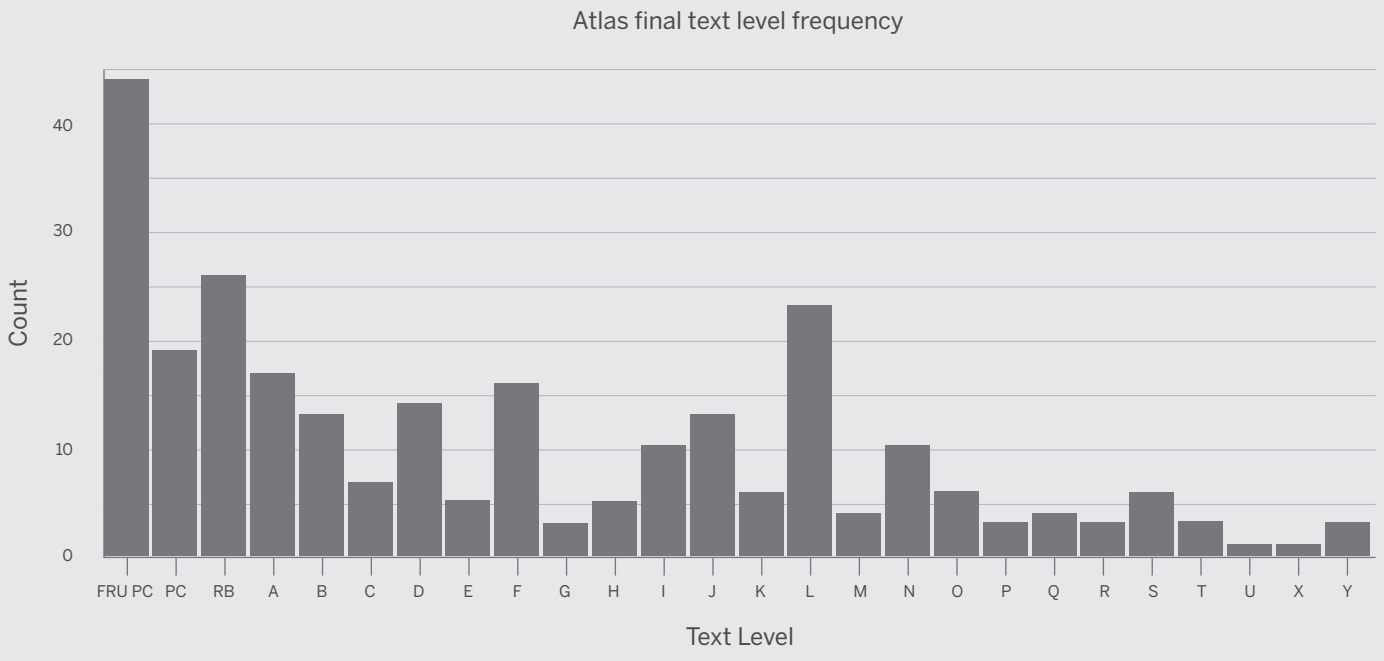
| Grade | Students (n) | Correlation With DIBELS Next Composite Score |
|---|---|---|
| All | 137 | 0.93 |
| 4 | 38 | 0.97 |
| 5 | 45 | 0.97 |
| 6 | 54 | 0.95 |

## Predictive Validity

Predictive validity provides an estimate of the extent to which student performance on the Atlas edition of TRC predicts scores on criterion measures administered at a later point in time, operationally defined as more than two months after the administration of Atlas. Estimated as the linear relationship between student performance on Atlas and the criterion measure, such predictive correlations are attenuated by time because students gain skills in the interim between testing occasions, and also by differences in the content specifications of the two assessments. Therefore, correlations between the assessment of interest (Atlas edition of TRC) and criterion measure (DIBELS Next composite score) greater than 0.7 are suggested to provide adequate criterion validity evidence (Kline, 2005). The relationship of final instructional reading level on the Atlas edition of TRC with the Composite score resulting from subsequent administration of DIBELS Next provides predictive validity evidence for Atlas.

Of 655 students administered the Atlas edition of TRC during Study A, final instructional reading levels were calculated for 281 students (Appendix 4); and 266 of those 281 students provided DIBELS composite scores at EOY 2013–2014. The resulting sample was composed of students from the following demographic categories: 48 percent female, 45 percent male, 7 percent unknown gender; 10

**Figure 8. Final Text Level Frequency for Students With DIBELS Composite Scores for Study A**



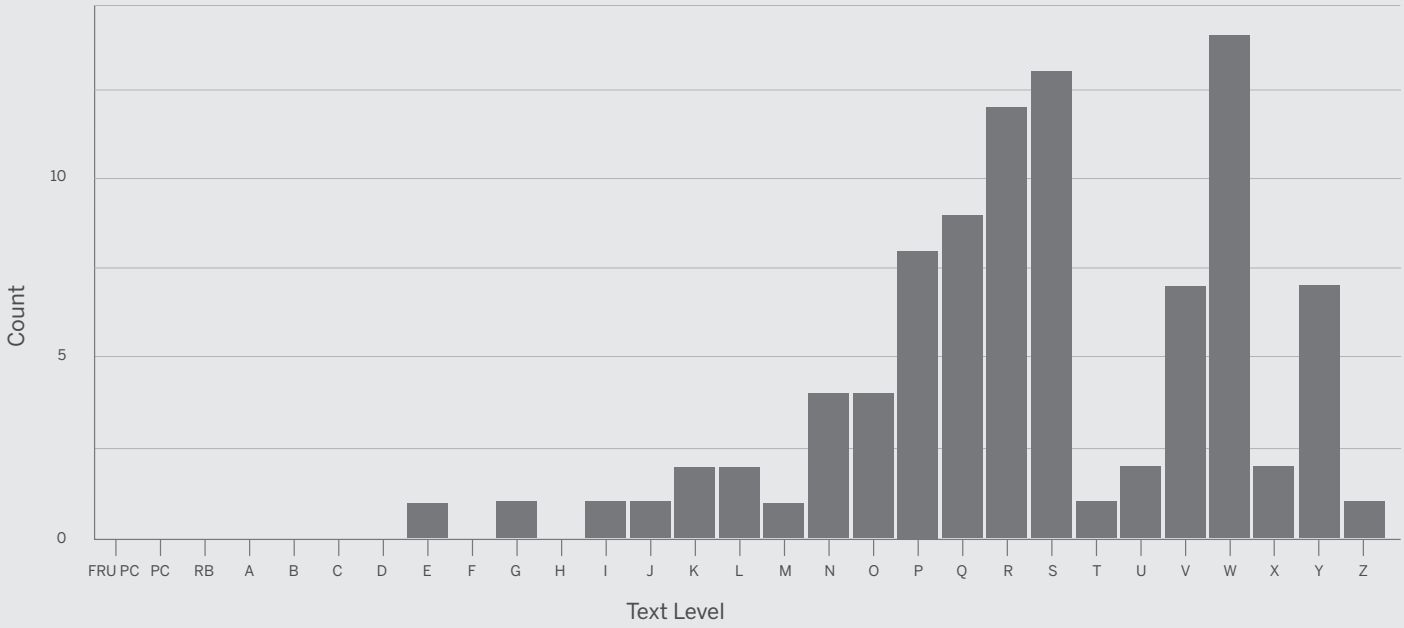
percent black, 33 percent white, 36 percent Hispanic, 21 percent other race or unknown race. Figure 8 presents the distributions of the final instructional reading levels for these 266 students.

**Table 25. Predictive Validity Evidence for Atlas in Kindergarten Through Grade 5 for Study A**

| Grade | Students (n) | Correlation With DIBELS Next Composite Score |
|---|---|---|
| All | 266 | 0.85 |
| K | 109 | 0.71 |
| 1 | 51 | 0.73 |
| 2 | 51 | 0.81 |
| 3 | 43 | 0.73 |
| 4 | 4 | 0.59 |
| 5 | 8 | 0.47 |

Of 434 students administered the Atlas edition of TRC during Study D, final instructional reading levels were calculated for 154 students (Appendix 4); 93 of those 154 students also provided DIBELS composite scores at EOY 2014–2015. The resulting sample was composed of students from the following demographic categories: 51 percent female, 49 percent male; 9 percent black, 65 percent white, 8 percent Hispanic, 19 percent other race or unknown race. Figure 9 presents the distributions of the final instructional reading levels for these 93 students.

Figure 9. Final Text Level Frequency for Students With DIBELS Composite Scores in Atlas 4–6 Study




Predictive validity evidence resulting from Study D is summarized in Table 26; the correlation between final instructional reading levels on Atlas and DIBELS Next composite scores ranged from 0.83 to 0.93, with an overall correlation of 0.81. The correlations are greater than 0.7, suggesting adequate evidence for the predictive validity of the Atlas edition of TRC with DIBELS Next composite score.

Table 26. Predictive Validity Evidence for Atlas in Grades 4 through 5 from Study D

| Grade | Students (n) | Correlation With DIBELS Next Composite Score |
|---|---|---|
| All | 93 | 0.81 |
| 4 | 27 | 0.93 |
| 5 | 34 | 0.86 |
| 6 | 32 | 0.83 |

## Construct Validity

Construct validity may be examined according to the developmental changes demonstrated by test performance for traits expected to increase with age (Anastasi, 1982). Reading is a skill that is expected to develop with age — as students read more, their reading comprehension skills and reading fluency improve, and, therefore, they can read more complex material. Atlas, as a reading assessment measure for kindergarten to Grade 6 students, provides a vertical scale that increases in student performance to reflect developmental changes in reading skills. Cross-sectional analysis of student performance at various ages collected at a single point in time is one method appropriate to provide construct validity evidence, presuming developmental changes or increasing student performance (Birren & Schaie, 2006). Descriptive statistics for final instructional text level on Atlas by grade were investigated to provide construct validity evidence.

Of the 655 students administered the Atlas edition of TRC during Study A, final instructional reading levels were calculated for 281 students (Appendix 4). The resulting sample was composed of 48 percent female, 45 percent male, 7 percent unknown gender; 10 percent black, 33 percent white, 36 percent Hispanic, 21 percent other race or unknown race. Figure 9 presents the distributions of the final instructional reading levels for the 281.

**Figure 10. Final Text Level Frequency for Study A**

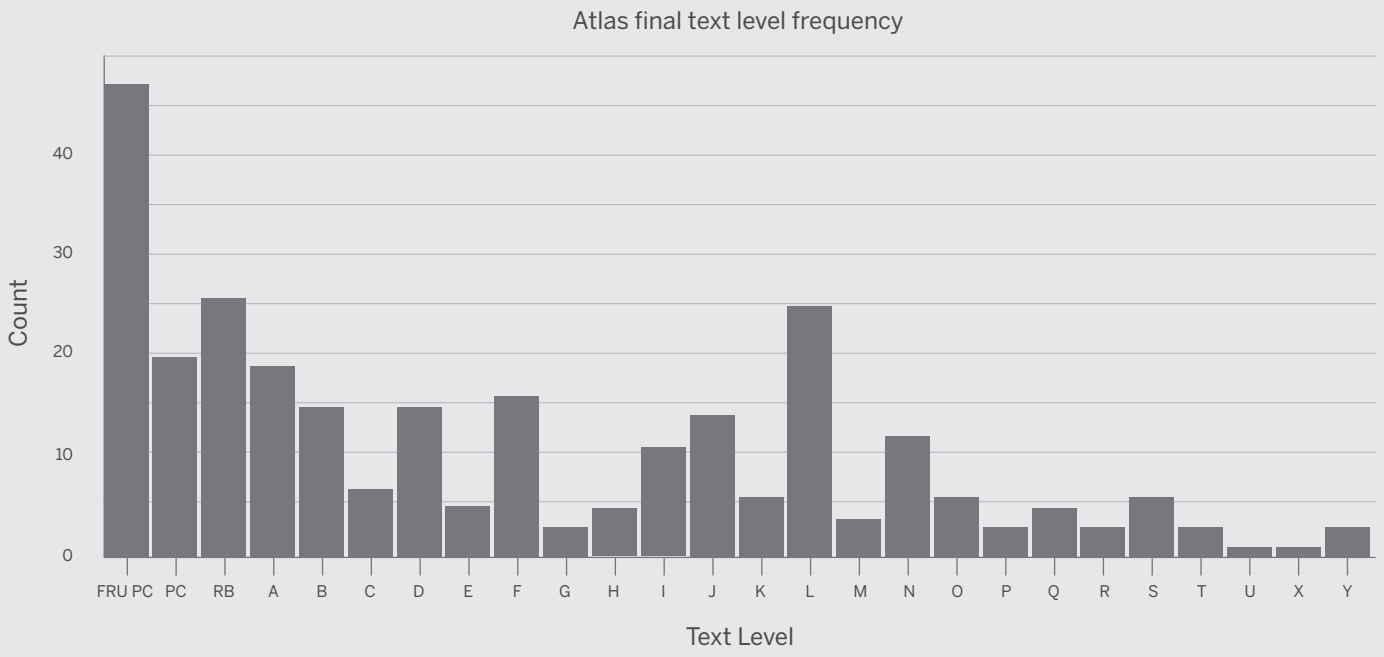


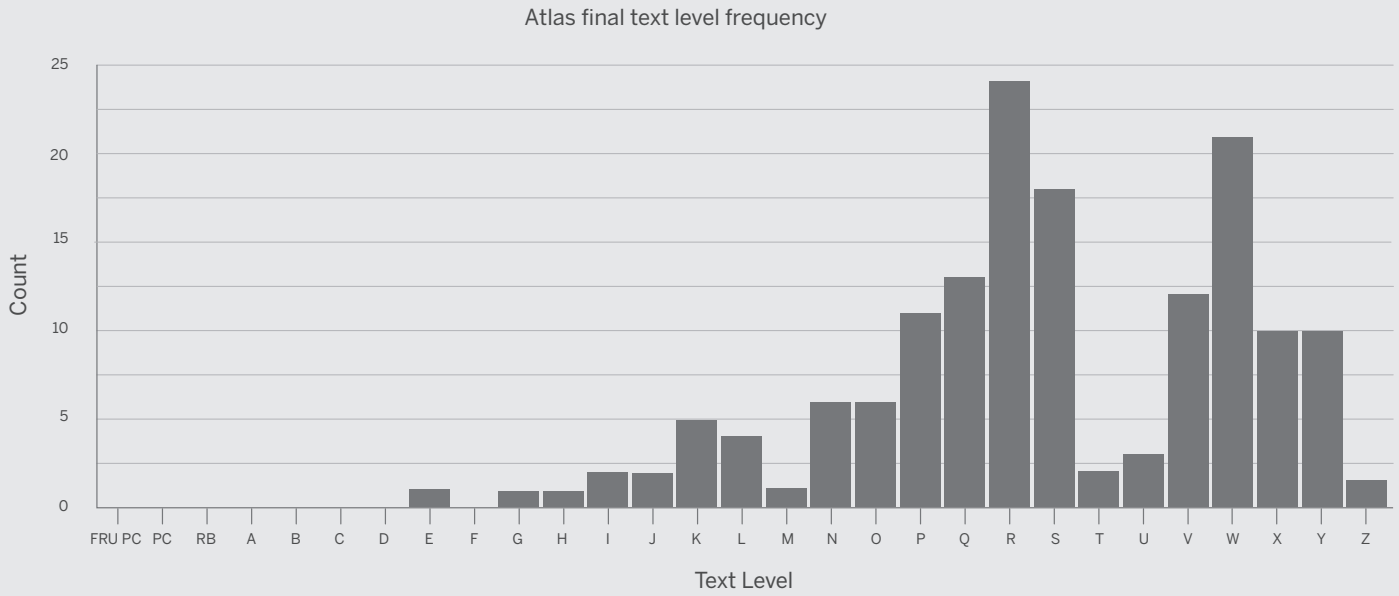
Table 27 provides descriptive statistics for the final instructional reading level resulting from administration of the Atlas edition of TRC. Final instructional reading levels are shown to increase with each grade, starting with RB/PC mean and median text levels at kindergarten, respectively, and increasing to U/T mean and median text levels at Grade 5. Further, student grade is shown to have a strong positive correlation with final instructional reading level (r = 0.83). These results are consistent with the assumption of increasing student performance (i.e., instructional reading level) according to development changes or grade level, providing evidence, therefore, for the construct validity of the Atlas edition of TRC.

**Table 27. Cross-Sectional Analysis of Student Performance for Study A**

| Grade | Students (n) | Text Level (Mean) | Text Level (Median) | SD |
|---|---|---|---|---|
| K | 113 | RB | PC | 2.50 |
| 1 | 55 | D | D | 4.14 |
| 2 | 58 | J | L | 4.12 |
| 3 | 43 | L | L | 5.20 |
| 4 | 4 | R | Q | 5.00 |
| 5 | 8 | U | T | 2.56 |

Additional construct validity evidence for the Atlas edition of TRC is provided by the 154 students in Study D for whom final instructional reading levels could be calculated. Figure 11 presents the distributions of the final instructional reading levels for these students.

## Figure 11. Final Text Level Frequency for Study D




Descriptive statistics for the final instructional reading level resulting from administration of the Atlas edition of TRC during Study D are provided in Table 28. Final instructional reading levels are shown to increase with each grade, starting with Q as the mean and median text level at Grade 4, increasing to text level T as the mean mean and median text level at Grade 6. These results are consistent with the assumption of increasing student performance (i.e., instructional reading level) according to developmental changes or grade level, providing further evidence for the construct validity of the Atlas edition of TRC.

Table 28. Cross-Sectional Analysis of Student Performance for Study D

| Grade | Students (n) | Text Level (Mean) | Text Level (Median) | SD |
|---|---|---|---|---|
| 4 | 43 | Q | Q | 4.54 |
| 5 | 51 | S | R | 3.85 |
| 6 | 60 | T | T | 4.10 |

## Benchmark Validation

### Standard Setting

Cut points for kindergarten through Grade 6 were established for the Atlas book set during a workshop convened April 12–13, 2014, in Brooklyn, New York. The Item Descriptor (ID) Matching method (Ferrara & Lewis, 2012), a standard-setting procedure appropriate for use with performance- based assessments that yield categorical results (e.g., Below Proficient, Proficient) such as TRC, was used to evaluate the Amplify Atlas book set against the CCSS for ELA to determine cut points for the four performance levels represented in TRC. In this procedure, participants with knowledge of and experience with the assessment and content area evaluate the tasks (i.e., the leveled texts and the associated comprehension activities) against content standards in an attempt to identify the tasks that best indicate the minimum expectations for each performance level, at each grade and time of year. The ID Matching method reduces cognitive burden on participants in comparison to judgmental activities required by other standard-setting methods and most clearly translates content standards into performance categories as compared to other methods of setting standards (Cizek, Bunch, & Koons, 2004).

A panel of 11 early reading and literacy experts — expert practitioners and researchers with a median of 17 years of experience and deep understanding of early reading instruction, curriculum, assessment, and development — was convened to determine the text levels which best indicated the cut points for the TRC performance levels ("Far Below Proficient," "Below Proficient," "Proficient," and "Above Proficient") when using the Atlas book set. The CCSS for ELA were used as the performance level standards to guide determinations of proficiency levels at each time of year (i.e., BOY, MOY, EOY) and grade. Panelists were supported in their decisions by the presentation of agreement and impact data and given the opportunity to react to this data and discuss revisions to cut points as necessary. Agreement data indicated the range and median of cut points provided by the panelists as a group; impact data described the percentage of students achieving at or above the panelist- selected cut points by grade and time of year. Impact data was simulated according to Item Response Theory (Bond & Fox, 2007), using student

performance on TRC from the 2012–2013 school year and statistical difficulty estimates for each Amplify Atlas book resulting from the field study to generate national student performance information appropriate for use in the ID Matching procedure. In total, panelists made decisions regarding 63 cut points: seven grades, three administration periods per grade, and three cut points per period. Full results are presented in Table 29. (See the TRC Standard Setting Research Report for more information on the standard-setting process and results.)

**Table 29. Text and Performance Levels for Amplify Atlas by Grade and Time of Year**

| Grade | TOY | Far Below | Below | Proficient | Above |
|---|---|---|---|---|---|
| K | BOY | < PC | PC | RB | A and above |
| | MOY | RB or below | A | B | C and above |
| | EOY | A or below | B | C to D | E and above |
| 1 | BOY | A or below | B | C to D | E and above |
| | MOY | C or below | D to E | F to G | H and above |
| | EOY | E or below | F to H | I | J and above |
| 2 | BOY | E or below | F to H | I | J and above |
| | MOY | H or below | I | J to K | L and above |
| | EOY | J or below | K | L to M | N and above |
| 3 | BOY | J or below | K | L to M | N and above |
| | MOY | K or below | L to M | N | O and above |
| | EOY | L or below | M to N | O to P | Q and above |
| 4 | BOY | L or below | M to N | O to P | Q and above |
| | MOY | N or below | O to P | Q | R and above |
| | EOY | P or below | Q | R to S | T and above |
| 5 | BOY | P or below | Q | R to S | T and above |
| | MOY | Q or below | R to S | T | U and above |
| | EOY | S or below | T | U to V | W and above |
| 6 | BOY | S or below | T | U to V | W and above |
| | MOY | U or below | V | W to X | Y and above |
| | EOY | V or below | W to X | Y to Z | * |

* No cut point set; no books available to classify students at this administration period and performance level.

## Additional Validity Evidence

A survey was designed to collect teacher feedback about TRC, Atlas, the Common Core State Standards, and participation in the Atlas field study. The survey was created using DatStat, a data management system, and administered via a link emailed to all potential respondents which included any educator who reviewed, administered and/or scored the Atlas edition of TRC within Study A. The survey link was distributed to participating educators in a study overview packet, with two additional completion reminders.

Survey results informed changes to the original test materials as necessary. Particular attention was given to feedback specific to texts or levels; this feedback was reviewed internally to determine whether a revision was warranted. Based on survey responses, revisions were made to book content, illustrations, comprehension questions, and/or administrative procedures as applicable.

The survey was composed of 41 questions divided into six thematic domains:

1. Demographics and Background Information
2. Common Core State Standards
3. mCLASS Reading3D – TRC
4. Amplify Atlas
5. Reading Records
6. The Atlas field study

The survey collected information about educators' use of, and familiarity and proficiency with TRC and reading records, teacher demographic information, perceptions of TRC, perceptions of Atlas, and perceptions of the CCSS and the alignment of CCSS with TRC and Atlas. There were 33 structured survey items (i.e., multiple choice), and all survey participants could provide open-ended feedback through six open-ended questions.

### Survey Results

The number of completed and consented responses totaled 46, representing eight districts and 21 schools.

**Educator Demographics.** Of the 46 respondents, all 46 reported reviewing the Atlas texts during the field study period with 23 administering Atlas as well. Most respondents had a master's degree or higher (68.9%, 31 respondents) and all survey respondents reported holding at least a bachelor's degree at minimum. Respondents most frequently reported teaching Grades K–2 (22 kindergarten, 24 Grade 1, 25

Grade 2) with approximately a quarter of respondents teaching Grades 4–6. Survey respondents demonstrated a median of 11 years of experience working in schools, with a median of 11 reported for years working at the elementary level and a median of three years working at her current school.

**Educator Use of TRC.** Most respondents reported advanced proficiency with TRC (56.8%), and a median of five years administering TRC. Twenty-one educators (47.7%) rated themselves as Proficient at administering running records and reported a median of 10 years experience. Most respondents were STEP and/or Mondo users (90.9%), with 52.3 percent using STEP and Mondo together, 36.4 percent using STEP only and 2.3 percent using Mondo only. Educators reported a median of 21 minutes to administer the TRC benchmark assessment and most self-rated as Advanced users of TRC (56.8%). Most respondents said they administer TRC due to district (97.5%) or school (92.5%) requirements, with fewer reporting administration of TRC due to state requirements (65.0%). Almost half of respondents said they use TRC for progress monitoring more than once per month (42.5%) and most reported using TRC for benchmark testing once per semester or benchmark period (67.5%). All respondents who use TRC for benchmarking reported conducting all (BOY, MOY, EOY) benchmark tests.

**Educator Perceptions of TRC.** Overall educator ratings of the TRC assessment were positive. Respondents were in the most agreement with the following statements:

- RC helps me select appropriate instructional reading materials and activities (90.9% agree or strongly agree)
- TRC helps me identify instructional goals for students (88.6% agree or strongly agree)
- TRC is useful to monitor the progress of students' reading ability (88.6% agree or strongly agree)

Respondents demonstrated the least agreement with the following statements:

- The amount of time required to administer TRC is appropriate, neither too long or too short (50.0% agree or strongly agree),
- TRC provides results that accurately represent a student's comprehension level (75.0% agree or strongly agree),
- TRC is a reliable assessment of reading ability (79.5% agree or strongly agree).

Additional responses support this concern for the amount of time it takes to administer TRC with one educator stating "some books are long in the higher levels" and another saying "it takes a long time to assess, especially if you aren't sure of a student's level."

Table 30. Perceptions of TRC

| Survey question | Strongly Agree | Agree | Disagree | Strongly Disagree | Not Applicable |
|---|---|---|---|---|---|
| Items demonstrating most agreement | | | | | |
| TRC helps me select appropriate instructional reading materials and activities | 27.30% | 63.60% | 2.30% | 2.30% | 4.50% |
| TRC helps me identify instructional goals for students | 29.50% | 59.10% | 4.50% | 2.30% | 4.50% |
| TRC is useful to monitor the progress of students' reading ability | 36.40% | 52.30% | 4.50% | 2.30% | 4.50% |
| Items demonstrating least agreement | | | | | |
| TRC is a reliable assessment of reading ability | 20.50% | 59.10% | 11.40% | 2.30% | 6.80% |
| TRC provides results that accurately represent a student's comprehension level | 20.50% | 54.50% | 18.20% | 2.30% | 4.50% |
| The amount of time required to administer TRC is appropriate, neither too long or too short | 4.50% | 45.50% | 36.40% | 9.10% | 4.50% |

Educator perceptions of Atlas. Most educators responding to the survey were in agreement that the Atlas book set appropriately measured students' oral reading accuracy and oral comprehension and that overall quality of the Atlas book set was the same as or higher than other existing book sets. Most respondents reported that student performance on Atlas was similar to student performance on other TRC book sets; however, the majority (70.5%) felt that the Atlas comprehension questions were more difficult than comprehension questions of other TRC book sets. Similar to the amount of time estimated to administer TRC (median of 21 minutes), educators reported a median of 20 minutes to administer the Atlas benchmark assessment.

Respondents were in the most agreement with the following statements about Atlas:

- The Atlas book set materials demonstrate sufficient coverage of cross-curricular subject areas (100% agree or strongly agree),
- The Atlas books provide interesting and compelling reading experiences for the students (100% agree or strongly agree), and
- The Atlas books do not demonstrate any cultural and gender bias or stereotype, do not showcase any violent or offensive material, and do not break any other social compliance codes (97.7% agree or strongly agree).

There was the least amount of agreement with the following statements:

- The modifications to Print Concepts in the Atlas book set allowed me to obtain important information about my students' reading skills (43.2% agree or strongly agree).
- The modifications to Reading Behaviors in the Atlas book set allowed me to obtain important information about my students' reading skills (43.2%).

However, a large number of respondents indicated that these statements were not applicable (40.9% and 43.2%, respectively). These responses imply that some educators did not administer Print Concepts or Reading Behaviors during the Atlas field trial, but those who did agreed that these measures allowed them to obtain important information about students' reading skills.

**Table 31. Educator Perceptions of Atlas**

| Survey question | Strongly Agree | Agree | Disagree | Strongly Disagree | Not Applicable |
|---|---|---|---|---|---|
| The Atlas book set materials demonstrate sufficient coverage of cross-curricular subject areas. | 25.00% | 75.00% | 0.00% | 0.00% | 0.00% |
| The Atlas passages provide interesting and compelling reading experiences for the students. | 22.70% | 77.30% | 0.00% | 0.00% | 0.00% |
| The Atlas passages provide students with reading experiences comparable to authentic reading activities. | 29.50% | 65.90% | 4.50% | 0.00% | 0.00% |
| The Atlas passages do not demonstrate any cultural and gender bias, stereotype or violence. | 25.00% | 72.70% | 2.30% | 0.00% | 0.00% |
| The Atlas passages provide effective graphical support. | 18.20% | 70.50% | 6.80% | 0.00% | 4.50% |
| Overall, the Atlas book set will assist me in providing reading instruction that is matched to student's instruction. | 25.00% | 68.20% | 2.30% | 0.00% | 4.50% |
| If provided the opportunity, I would prefer to use the Atlas book set over my existing book set. | 38.60% | 38.60% | 13.60% | 0.00% | 9.10% |
| The modifications to Print Concepts in the Atlas book set allowed me to obtain important information about my students' reading skills. | 11.40% | 31.80% | 13.60% | 2.30% | 40.90% |
| The modifications to Reading Behaviors in the Atlas book set allowed me to obtain important information about my students' reading skills. | 11.40% | 31.80% | 11.40% | 2.30% | 43.20% |

| | | | | | |
|---|---|---|---|---|---|
| Eliminating the reread of the text prior to completing comprehension tasks (retell and/or oral comprehension) is an improvement over previous versions of TRC. | 15.90% | 50.00% | 4.50% | 2.30% | 27.30% |
| Making Written Comprehension optional within the Atlas book set is an improvement over previous versions of TRC. | 25.00% | 40.90% | 13.60% | 4.50% | 15.90% |

In addition to rating features of the Atlas book set, educators were also asked to identify up to three specific books to which they would most like to see changes made. Most of the identified books were lower-level books with suggestions for changes to graphical support and decreasing the difficulty of the text and/or comprehension questions. The most commonly cited books were:

- Bugs (Level A, 6 responses)
- Hands Can Do a Lot (Level A, 4 responses)
- Fun With Clay (Level D, 3 responses)
- A Pumpkin Grows (Level D, 3 responses)
- I Can Help (Level E, 3 responses)

Suggested changes were reviewed internally within the context of student performance data from the field trial, and necessary edits were made to books, illustrations, and comprehension questions.

Common Core State Standards and TRC. Overall, educators reported familiarity with the CCSS for ELA and rated Atlas as well aligned to the grade-level expectations of CCSS. Twenty-one educators (47.7%) rated themselves as Proficient with CCSS, but most frequently identified their district's implementation as Basic (40.9%). Respondents reported an average of 2.9 hours of professional development related to CCSS implementation. All respondents agreed or strongly agreed that Atlas reflects the CCSS with respect to vocabulary, sentence structure, organization, and the Atlas book set materials reflect the CCSS with regard to appropriate coverage of literary (fiction) and informational texts. Most respondents also agreed with the following statements:

- Overall, the Atlas book set will assist me in providing reading instruction that is matched to student's instruction (93.2% agree or strongly agree).
- Overall the Atlas book set will assist me in addressing instructional shifts called for by the CCSS (93.2% agree or strongly agree).

The results suggest that educators think Atlas is an assessment with a strong base in the expectations of the CCSS.

Table 32. CCSS and Atlas

| Survey question | Strongly Agree | Agree | Disagree | Strongly Disagree | Not Applicable |
|---|---|---|---|---|---|
| The Atlas book set materials reflect the CCSS with respect to vocabulary, sentence structure, organization. | 13.60% | 86.40% | 0.00% | 0.00% | 0.00% |
| The Atlas book set materials reflect the CCSS with regard to appropriate coverage of literary (fiction) and informational (nonfiction) texts. | 27.30% | 72.70% | 0.00% | 0.00% | 0.00% |
| Overall, the Atlas book set will assist me in providing reading instruction that is matched to student's instruction. | 25.00% | 68.20% | 4.50% | 0.00% | 2.30% |
| Overall the Atlas book set will assist me in addressing instructional shifts called for by the CCSS. | 15.90% | 77.30% | 4.50% | 0.00% | 2.30% |

**Summary.** Overall, field study participants indicated satisfaction with the quality and content of the Atlas books, illustrations, oral comprehension questions and other assessment features. Respondents also agreed that Atlas demonstrated alignment with the CCSS for ELA. Survey results also suggest that Atlas texts are similar difficulty levels to other existing TRC book sets with the exception of the more difficult oral comprehension questions. Respondents indicated areas for improvement specific to particular books (e.g., Bugs) or specific text levels (e.g., Print Concepts). Most requests for changes were regarding lower-level books (A–E) which may be attributed to the necessity of fine-tuned differences among texts at these earlier developmental stages.

Suggestions and comments from field study participants were given careful consideration within the context of student performance data, and revisions were made to the Atlas materials as necessary. Edits to Atlas materials were made when data from both student performance and qualitative feedback converged and when qualitative feedback strongly suggested a need for an edit. Revisions included changes to book content, illustrations, oral comprehension questions, and administrative procedures.

# References


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

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

Anastasi, A. (1982). **Psychological testing** (5th ed.). New York: McMillan Publishing.

Bean, R., Cassidy, J., & Grumet, J. (2002). What do reading specialists do? Results from a national survey. **The Reading Teacher**, 55(8), 736–744.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. **Journal of Learning Disabilities**, 35(1), 39–56.

Birren, J. E., & Schaie, K. W. (Eds.). (2006). **Handbook of the psychology of aging** (6th ed.). San Diego, CA: Elsevier.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. **Phi Delta Kappan**, 80(2), 139–148.

Bond, T. G., & Fox., C. M. (2007). Applying the Rasch Model: **Fundamental measurement in the human sciences** (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Carver, R. (1990). **Reading rate: A review of research and theory**. San Diego: Academic Press.

Chall, J. S. (1983). **Stages of reading development**. New York: McGraw-Hill.

Chall, J. S. (1996). **Learning to read: The great debate (revised, with a new forward)**. New York: McGraw-Hill.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. **Psychological Assessment**, 6(4), 284.

Cizek, G. J., & Bunch, M. B. (2007). **Standard setting: A guide to establishing and evaluating performance standards on tests**. New York: SAGE Publications Ltd.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). **Setting performance standards: Contemporary methods. Educational Measurement: Issues and Practice**, 23(4), 31–50.

Clay, M. M. (1991). **Becoming literate: The construction of inner control**. Portsmouth, NH: Heinemann.

Clay, M. M. (1993a). **An observation survey of early literacy achievement**. Portsmouth, NH:Heinemann.

Clay, M. M. (1993b). **Reading recovery: A guidebook for teachers in training**. Portsmouth, NH: Heinemann.

Clay, M. M. (1998). **By different paths to common outcomes**. York, ME: Stenhouse Publishers.

Clay, M. M. (2001). **Change over time in children's literacy development**. Portsmouth, NH:Heinemann.

Clay, M. M. (2002). **An observation survey of early literacy achievement** (2nd ed.). Portsmouth, NH: Heinemann.

Clay, M. M. (2005). **An observation survey of early literacy achievement** (3rd ed.). Portsmouth, NH: Heinemann.

Cox, B. E. & Hopkins, C. J. (2006). Building on theoretical principles gleaned from Reading Recovery to inform classroom practice. **Reading Research Quarterly**, 41(2), 254–267.

Dowhower, S. L. (1987). Effects of repeated reading on second-grade transitional readers' fluency and comprehension. **Reading Research Quarterly**, 22(4), 389–340.

Duke, N. K. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. **Reading Research Quarterly**, 35, 202–224.

Dynamic Measurement Group. (2010). DIBELS Next Benchmark Goals and Composite Scores. Retrieved from https://dibels.uoregon.edu/docs/DIBELSNextFormerBenchmarkGoals.pdf

Dynamic Measurement Group. (2010). Dynamic Indicators of Basic Early Literacy Skills (7th ed.).Eugene, OR: Author.

Ehri, L. C. (1991). Development of the ability to read words. In R. Barr, M. L. Kamil, P. Mosenthal, and P. D. Pearson (Eds.), **Handbook of Reading Research** (Vol. 2, pp. 383–417). White Plains, NY: Longman.

Ehri, L. C. (1995). Phases of development in learning to read words by sight. **Journal of Research in Reading**, 18, 116–125.

Embretson, S. E., & Reise, S. P. (2000). **Item response theory for psychologists**. New York: Psychology Press.

Ferrara, S., & Lewis, D. (2012). The item-descriptor matching method. **Setting performance standards** (2nd ed., pp. 255–282). New York, NY: Routledge.

Fleiss, J.L. (1981). **Statistical methods for rates and proportions** (2nd ed.). New York: John Wiley.

Fountas, I.C., & Pinnell, G.S. (1996). **Guided reading: Good first teaching for all children**. Portsmouth, NH: Heinemann.

Fountas, I. C., & Pinnell, G. S. (1999). **Matching books to readers: Using leveled books in guided reading, K–3**. Portsmouth, NH: Heinemann.

Fountas, I.C., & Pinnell, G.S. (2011). **The continuum of literacy learning, grades PreK–8**. Portsmouth, NH: Heinemann.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. **School Psychology Review**, 21(1), 45–58.

Fuchs, L. S., Fuchs, D., & Deno, S. L. (1982). Reliability and validity of curriculum-based informal reading inventories. Reading Research Quarterly, 18(1), 6–26.

Garrison, C., & Ehringhaus, M. (2007). Formative and summative assessments in the classroom. Retrieved from http://www.nmsa.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx

Good, R. H., Kaminski, R., Dewey, E., Walin, J., Powell-Smith, K., & Latimer, R. (2013). **DIBELS Next Technical Manual**, Eugene, OR: Dynamic Measurement Group, Inc.

Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. **Harvard Educational Review**, 81(4), 710–744.

Guskey, T. R. (2003). How classroom assessments improve learning. **Educational Leadership**, 60(5), 6–11.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. **Tutorials in quantitative methods for psychology**, 8(1), 23.

Hoffman, J.V., McCarthey, S.J., Abbott, J., Christian, C., Corman, L., Curry, C., Dressman, M., Elliott, B., Matherne, D., & Stahle, D. (1994). So what's new in the new basals? A focus on first grade. **Journal of Reading Behavior**, 26, 47–73.

Holmes, J. and Singer, H. (1961). The substrata-factor theory: Substrata-factor differences underlying reading ability in known groups. US Office of Education, Final Report No. 538, SAE 8176.

Johns, J. L. (1980). First graders' concepts about print. **Reading Research Quarterly**, 15(4), 529–549.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. **Journal of Educational Psychology**, 8(4), 437–44

Justice, L. M., & Ezell, H. K. (2001). Word and print awareness in 4-year-old children. **Child Language Teaching and Therapy**, 17(3), 207–22.

Kamberelis, G. (1998). Relations between children's literacy diets and genre development: You write what you read. **Literacy Teaching and Learning**, 3, 7–53.

Kamil, M. L., & Lane, D. (1997). **A classroom study of the efficacy of using information text for first- grade reading instruction.** Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kaminski, R., & Good, R. (1996). Toward a technology for assessing basic early literacy skills. **School Psychology Review**, 25(2), 215–227.

Kane, M. (2001). Current Concerns In Validity Theory. **Journal of Educational Measurement**, 38(4), 319–342.

Kintsch, W., & Dijk, T. A. Van. (1978). Toward a model of text comprehension and production.**Psychological Review**, 85(5), 363–394.

Kline, R. B. (2005). **Principles and Practice of Structural Equation Modeling** (2nd ed.). New York: Guilford Press.

Linacre, J. M. (2014). **Winsteps® Rasch measurement computer program User's Guide**. Beaverton, OR: Winsteps.com

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. **Journal of School Psychology**, 47(5), 315–335.

Masters, G. N. (1982). A Rasch model for partial credit scoring. **Psychometrika**, 47, 149–174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), **Educational measurement** (3rd ed.) (pp. 13–103). New York: Macmillan.

National Early Literacy Panel. (2008). **Developing early literacy: Report of the national early literacy panel**. Washington, DC: National Institute for Literacy.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). Common Core State Standards English Language Arts. Washington, DC: Author.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995).Listening to children read aloud. Washington, DC: U.S. Government Printing Office.

Pressley, M., Wharton-McDonald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., Baker, K., Brooks, G., Cronin, J., Nelson, E., & Woo, D. (2001). A study of effective first-grade literacy instruction. **Scientific Studies of Reading**, 5(1), 35–58.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. **School Psychology Quarterly**, 20(3), 304–317.

Ross, J. A. (2004). Effects of running records assessment on early literacy achievement: Results of a controlled experiment. **Journal of Educational Research**, 97(4), 186–194.

Rumelhart, D. E. (1977). Toward an interactive model of reading. **Attention and performance**, 6, 573– 603.

Rumelhart, D. E. (1994). Toward and interactive model of reading. In R. B. Ruddell, M. R. Ruddell, and H. Singer (Eds.), Theoretical Models and Processes of Reading (4th Edition), 864–894, Newark, DE: International Reading Association.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2013). **Assessment in special and inclusive education**. Cengage Learning.

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

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. **Journal of Educational Measurement**, 28(3), 234–247.

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. **Psychological bulletin**, 86(2), 420.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. **Reading Research Quarterly**, 21(4), 360–407.

Stevens, S. S. (1946). On the theory of scales of measurement. **Science**, 103(2684), 677–680.

Stiggins, R. (2002). Assessment crisis: The absence of assessment FOR learning. **Phi Delta Kappan**, 83(10), 758–765.

Stuart, M. (1995). Prediction and qualitative assessment of five- and six-year-old children's reading: A longitudinal study. **British Journal of Educational Psychology**, 65, 287–296.

U.S. Department of Education. (2012). Elementary/ Secondary Information System. Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://nces.ed.gov/ccd/elsi

Von Secker, C., Zhao, H., & Powell, M. (2008). **Attainment of end-of-year reading benchmarks in kindergarten to grade 2**. Rockville, MD: Montgomery County Public School.

Vygotsky, L. S. (1978). **Mind in society: The development of higher psychological processes**.Cambridge, MA: Harvard University Press.

Yopp, R.H., & Yopp, H.K. (2000). Sharing informational text with young children. **The Reading Teacher**, 53, 410–423.

Zhao, H. & Von Secker, C. (2008). **Evaluation of the criterion-related validity of Montgomery County Public Schools Assessment Program in Primary Reading**. Rockville, MD: Montgomery County Public Schools.

# Appendix 1. Demographic Comparison of National Schools, TRC Schools, and Atlas Field Study Schools

**Table 33. Demographic Comparison Table**

| | Schools Nationwide | | | | | | | Schools Using TRC in 2013–2014 | | | | | | | Atlas Field Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Overall |
| Sample Size (n) | | | | | | | | | | | | | | | |
| States | 51 | 51 | 51 | 51 | 51 | 51 | 51 | 19 | 20 | 20 | 20 | 16 | 14 | 20 | 9 |
| Districts | 14,556 | 14,606 | 14,600 | 14,585 | 14,575 | 14,632 | 14,904 | 358 | 368 | 363 | 228 | 118 | 96 | 370 | 7 |
| Schools | 50,884 | 51,540 | 51,547 | 51,508 | 51,171 | 49,818 | 58,500 | 1,978 | 2,029 | 2,004 | 1,383 | 509 | 417 | 2,117 | 29 |
| Educators | | | | | | | | 13,906 | 14,669 | 14,337 | 9,628 | 2,852 | 2,227 | 57,619 | 47 |
| Students | 20,607,036 | 21,036,776 | 20,785,859 | 20,739,751 | 20,310,843 | 19,143,629 | 122,623,894 | 637,162 | 652,941 | 623,640 | 446,072 | 79,367 | 70,370 | 2,509,552 | 654 |
| Geographic Region (%) | | | | | | | | | | | | | | | |
| Midwest | 24.45 | 24.62 | 24.57 | 24.64 | 24.59 | 24.17 | 24.51 | 31.78 | 31.92 | 31.72 | 11.21 | 17.49 | 17.03 | 26.77 | 19.75 |
| Northeast | 15.80 | 15.91 | 15.88 | 15.78 | 15.55 | 15.18 | 15.69 | 0.81 | 0.99 | 1.05 | 1.37 | 1.96 | 2.16 | 1.14 | 0.00 |
| South | 34.30 | 34.12 | 34.19 | 34.20 | 34.31 | 34.31 | 34.24 | 65.39 | 64.48 | 64.64 | 84.24 | 74.07 | 74.58 | 69.11 | 14.81 |
| West | 25.44 | 25.35 | 25.36 | 25.37 | 25.55 | 26.34 | 25.57 | 2.02 | 2.61 | 2.59 | 3.18 | 6.48 | 6.24 | 2.98 | 65.43 |

| | Schools Nationwide | | | | | | | Schools Using TRC in 2013–2014 | | | | | | | Atlas Field Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Overall |
| Location Relative to Population Centers (%) | | | | | | | | | | | | | | | |
| City: large | 15.53 | 15.52 | 15.31 | 15.26 | 15.25 | 15.71 | 15.43 | 10.26 | 10.49 | 10.47 | 10.77 | 19.45 | 21.58 | 11.58 | 30.86 |
| City: mid-size | 6.45 | 6.40 | 6.43 | 6.42 | 6.46 | 6.67 | 6.47 | 7.23 | 7.29 | 7.13 | 8.32 | 9.43 | 9.35 | 7.64 | 34.57 |
| City: small | 7.57 | 7.51 | 7.52 | 7.51 | 7.52 | 7.55 | 7.53 | 8.94 | 8.87 | 9.13 | 7.01 | 7.66 | 6.95 | 8.47 | 7.41 |
| Suburb: large | 24.31 | 24.50 | 24.56 | 24.53 | 24.48 | 24.08 | 24.41 | 12.83 | 13.30 | 13.07 | 7.16 | 6.88 | 5.76 | 11.34 | 11.11 |
| Suburb: mid-size | 2.86 | 2.86 | 2.86 | 2.87 | 2.83 | 2.76 | 2.84 | 4.09 | 4.04 | 4.09 | 5.35 | 6.88 | 6.24 | 4.57 | 0.00 |
| Suburb: small | 1.86 | 1.89 | 1.88 | 1.87 | 1.86 | 1.85 | 1.87 | 1.21 | 1.23 | 1.15 | 1.23 | 0.59 | 0.72 | 1.14 | 0.00 |
| Town: fringe | 1.53 | 1.52 | 1.54 | 1.54 | 1.53 | 1.48 | 1.52 | 1.47 | 1.48 | 1.45 | 1.37 | 0.59 | 0.72 | 1.36 | 0.00 |
| Town: distant | 5.35 | 5.39 | 5.41 | 5.38 | 5.34 | 5.27 | 5.36 | 8.29 | 8.52 | 8.43 | 8.68 | 7.27 | 7.91 | 8.36 | 0.00 |
| Town: remote | 3.68 | 3.73 | 3.72 | 3.74 | 3.71 | 3.53 | 3.68 | 0.91 | 0.89 | 0.90 | 0.80 | 0.20 | 0.00 | 0.79 | 0.00 |
| Rural: fringe | 12.65 | 12.65 | 12.71 | 12.81 | 12.85 | 12.76 | 12.74 | 21.93 | 21.58 | 21.65 | 24.44 | 20.83 | 19.90 | 22.02 | 14.81 |
| Rural: distant | 11.55 | 11.46 | 11.45 | 11.48 | 11.53 | 11.62 | 11.52 | 20.46 | 19.95 | 20.20 | 21.69 | 16.70 | 16.79 | 20.06 | 1.23 |
| Rural: remote | 6.67 | 6.59 | 6.62 | 6.59 | 6.63 | 6.73 | 6.64 | 2.37 | 2.36 | 2.34 | 3.18 | 3.54 | 4.08 | 2.66 | 0.00 |
| Missing | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| School Type and Characteristics (%) | | | | | | | | | | | | | | | |
| Schoolwide Title I | 58.28 | 58.24 | 58.13 | 58.06 | 58.00 | 57.77 | 58.08 | 74.28 | 74.19 | 74.06 | 80.48 | 83.89 | 85.85 | 76.40 | 45.68 |
| Charter school | 6.05 | 6.03 | 5.97 | 5.92 | 5.86 | 6.21 | 6.01 | 2.07 | 2.07 | 2.09 | 2.60 | 2.95 | 2.88 | 2.26 | 0.00 |

| | Schools Nationwide | | | | | | | Schools Using TRC in 2013–2014 | | | | | | | Atlas Field Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Overall |
| Regular school | 97.92 | 97.65 | 97.51 | 97.39 | 97.22 | 96.87 | 97.43 | 99.70 | 99.70 | 99.7 | 99.57 | 99.41 | 99.76 | 99.66 | 100.00 |
| Special education school | 1.13 | 1.23 | 1.33 | 1.40 | 1.45 | 1.57 | 1.35 | 0.15 | 0.05 | 0.05 | 0.07 | 0.20 | 0.24 | 0.10 | 0.00 |
| Vocational school | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Alternative/ other school | 0.94 | 1.11 | 1.14 | 1.20 | 1.31 | 1.54 | 1.20 | 0.15 | 0.25 | 0.25 | 0.36 | 0.39 | 0.00 | 0.24 | 0.00 |
| Pupil to teacher ratio | 16.77 | 16.71 | 16.68 | 16.69 | 16.67 | 16.74 | 16.71 | 15.90 | 15.93 | 15.94 | 15.8 | 16.19 | 16.2 | 15.93 | 18.89 |
| Percentage of free/ reduced lunch | 53.02 | 53.02 | 52.98 | 52.95 | 53.00 | 53.13 | 53.02 | 57.1 | 57.09 | 57.01 | 59.42 | 61.76 | 63.66 | 58.07 | 32.38 |
| Student Characteristics (%) | | | | | | | | | | | | | | | |
| Male | 51.69 | 51.73 | 51.77 | 51.78 | 51.81 | 51.87 | 51.77 | 51.75 | 51.76 | 51.75 | 51.77 | 51.71 | 51.66 | 51.75 | 51.01 |
| Female | 48.31 | 48.27 | 48.23 | 48.22 | 48.19 | 48.13 | 48.23 | 48.25 | 48.24 | 48.25 | 48.23 | 48.29 | 48.34 | 48.25 | 48.99 |
| White | 53.29 | 53.30 | 53.44 | 53.49 | 53.45 | 52.99 | 53.33 | 57.19 | 56.96 | 57.13 | 53.54 | 51.38 | 48.06 | 55.70 | 44.31 |
| Black | 15.55 | 15.59 | 15.54 | 15.56 | 15.60 | 15.89 | 15.62 | 19.99 | 19.86 | 19.76 | 23.06 | 24.00 | 26.36 | 20.98 | 8.99 |
| Hispanic | 22.16 | 22.15 | 22.05 | 21.99 | 21.99 | 22.13 | 22.08 | 14.65 | 15.02 | 14.98 | 15.78 | 17.64 | 18.48 | 15.38 | 26.73 |
| Am. Indian/ Alaskan Native | 1.71 | 1.70 | 1.71 | 1.71 | 1.74 | 1.75 | 1.72 | 1.09 | 1.08 | 1.08 | 1.44 | 0.40 | 0.43 | 1.07 | 0.21 |
| Asian | 3.98 | 3.96 | 3.95 | 3.96 | 3.95 | 3.97 | 3.96 | 2.82 | 2.84 | 2.82 | 2.16 | 2.64 | 2.71 | 2.70 | 14.76 |

| | Schools Nationwide | | | | | | | Schools Using TRC in 2013–2014 | | | | | | | Atlas Field Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Kindergarten | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 | Overall | Overall |
| Hawaiian Native/ Pacific Islander | 0.39 | 0.39 | 0.39 | 0.39 | 0.39 | 0.40 | 0.39 | 0.11 | 0.11 | 0.11 | 0.11 | 0.12 | 0.12 | 0.11 | 0.30 |
| Two or more races | 2.92 | 2.91 | 2.92 | 2.90 | 2.88 | 2.87 | 2.90 | 4.13 | 4.14 | 4.11 | 3.91 | 3.83 | 3.84 | 4.06 | 4.70 |
| Student Character-istics (%) | ... on NAEP Reading (2013) | | | | | | | ... on TRC | | | | | | | |
| Far Below | × | × | × | × | N/A | × | | 64.03 | 33.66 | 33.76 | 33.49 | 39.94 | 42.06 | 42.04 | 32.94 |
| Below | × | × | × | × | 0.68 | × | | 10.48 | 16.29 | 19.71 | 17.97 | 17.62 | 16.81 | 15.96 | 20.07 |
| Proficient | × | × | × | × | 0.35 | × | | 14.17 | 20.79 | 22.97 | 17.51 | 19.03 | 25.49 | 19.15 | 22.41 |
| Above | × | × | × | × | N/A | × | | 11.32 | 29.26 | 23.56 | 31.03 | 23.42 | 15.64 | 22.85 | 24.58 |

# Appendix 2. Book Difficulty

**Table 34. Overall Book Performance Difficulty Estimates**

| Text Level | Mean Difficulty | Minimum Difficulty | Maximum Difficulty |
|---|---|---|---|
| A | –26.90 | –27.78 | –26.25 |
| B | –22.06 | –22.56 | –21.42 |
| C | –16.57 | –16.71 | –16.37 |
| D | –12.53 | –12.92 | –12.02 |
| E | –9.54 | –9.97 | –9.07 |
| F | –7.03 | –7.53 | –6.56 |
| G | –3.86 | –4.43 | –3.37 |
| H | –1.89 | –2.63 | –1.37 |
| I | 0.80 | 0.45 | 1.24 |
| J | 3.52 | 2.47 | 4.21 |
| K | 6.33 | 6.07 | 6.68 |
| L | 9.79 | 9.16 | 10.27 |
| M | 13.77 | 13.33 | 14.24 |
| N | 16.83 | 16.10 | 17.87 |
| O | 20.47 | 19.93 | 21.22 |
| P | 23.16 | 22.58 | 23.74 |
| Q | 26.79 | 26.42 | 27.53 |
| R | 29.71 | 29.40 | 29.93 |
| S | 32.87 | 32.32 | 33.62 |
| T | 36.46 | 35.65 | 37.24 |
| U | 40.23 | 39.98 | 40.36 |
| V | 43.89 | 43.63 | 44.14 |
| W | 47.74 | 47.11 | 48.37 |
| X | 53.69 | 53.43 | 53.95 |
| Y | 58.72 | 57.94 | 59.50 |
| Z | 64.76 | 64.42 | 65.10 |

**Note:** There are two to three books per text level.

Table 35. Reading Record Accuracy Difficulty Estimates

| Text Level | Mean Difficulty | Minimum Difficulty | Maximum Difficulty |
|---|---|---|---|
| A | –30.18 | –32.67 | –28.12 |
| B | –23.38 | –23.80 | –22.95 |
| C | –12.76 | –14.48 | –11.67 |
| D | –6.22 | –6.62 | –6.01 |
| E | –0.77 | –2.86 | 0.65 |
| F | 5.22 | 2.90 | 7.30 |
| G | 8.01 | 6.86 | 8.66 |
| H | 10.93 | 9.64 | 11.88 |
| I | 14.05 | 13.68 | 14.45 |
| J | 19.55 | 17.99 | 21.05 |
| K | 23.83 | 23.17 | 24.41 |
| L | 29.00 | 27.27 | 30.68 |
| M | 31.82 | 29.72 | 33.73 |
| N | 36.89 | 35.10 | 38.07 |
| O | 41.93 | 40.63 | 43.60 |
| P | 43.68 | 42.25 | 44.65 |
| Q | 49.09 | 46.92 | 50.69 |
| R | 50.96 | 48.95 | 53.70 |
| S | 52.31 | 50.07 | 55.92 |
| T | 55.13 | 53.87 | 56.31 |
| U | 59.17 | 55.92 | 61.47 |
| V | 63.05 | 61.82 | 64.28 |
| W | 63.53 | 62.86 | 64.20 |
| X | 68.43 | 67.36 | 69.50 |
| Y | 72.48 | 71.78 | 73.17 |
| Z | 71.69 | 67.55 | 75.83 |

**Note:** There are two to three books per text level.

Table 36. Retell/Recall Difficulty Estimates

| Text Level | Mean Difficulty | Minimum Difficulty | Maximum Difficulty |
|---|---|---|---|
| A | –9.60 | –10.57 | –8.89 |
| B | –4.27 | –5.43 | –2.95 |
| C | 2.19 | 1.44 | 3.59 |
| D | 6.27 | 5.76 | 6.67 |
| E | 10.77 | 9.91 | 11.57 |

**Note:** There are two to three books per text level.

Table 37. Oral Comprehension Difficulty Estimates

| Text Level | Mean Difficulty | Minimum Difficulty | Maximum Difficulty |
|---|---|---|---|
| D | –56.29 | –57.82 | –54.49 |
| E | –49.27 | –49.89 | –48.48 |
| F | –42.21 | –42.71 | –41.68 |
| G | –33.95 | –34.79 | –33.12 |
| H | –28.24 | –30.02 | –26.98 |
| I | –21.21 | –21.97 | –20.47 |
| J | –14.46 | –15.96 | –13.30 |
| K | –7.60 | –7.99 | –7.01 |
| L | –0.05 | –0.74 | 0.43 |
| M | 8.02 | 7.55 | 8.51 |
| N | 14.61 | 13.76 | 15.90 |
| O | 21.72 | 20.90 | 22.76 |
| P | 27.54 | 26.74 | 28.33 |
| Q | 34.26 | 33.77 | 35.18 |
| R | 39.97 | 39.48 | 40.32 |
| S | 45.76 | 45.12 | 46.57 |
| T | 51.89 | 50.88 | 52.86 |
| U | 57.96 | 57.73 | 58.09 |
| V | 63.72 | 63.43 | 64.01 |
| W | 69.47 | 68.71 | 70.23 |
| X | 77.49 | 77.16 | 77.82 |
| Y | 84.22 | 83.33 | 85.11 |
| Z | 91.70 | 91.28 | 92.12 |

**Note:** There are two to three books per text level.

# Appendix 3. Item Statistics for PC and RB Books

Table 38. Item Statistics for PC Book 1

| Item # | Difficulty Estimation | Standard Error of Difficulty | Infit | Outfit | Point-Biserial Correlations |
|---|---|---|---|---|---|
| 1 | –2.83 | 0.75 | 0.94 | 0.44 | 0.30 |
| 2 | –2.83 | 0.75 | 0.79 | 0.31 | 0.42 |
| 3 | –0.54 | 0.44 | 1.18 | 1.10 | 0.21 |
| 4 | –2.35 | 0.63 | 0.95 | 2.22 | 0.21 |
| 5 | –2.83 | 0.75 | 0.79 | 0.31 | 0.42 |
| 6 | –2.83 | 0.75 | 0.90 | 0.50 | 0.30 |
| 7 | –2.83 | 0.75 | 1.02 | 0.72 | 0.18 |
| 8 | –0.95 | 0.46 | 1.41 | 1.28 | 0.02 |
| 9 | –1.17 | 0.48 | 0.79 | 1.01 | 0.50 |
| 10 | –0.95 | 0.46 | 1.23 | 1.39 | 0.14 |
| 11 | –0.34 | 0.43 | 0.71 | 0.76 | 0.63 |
| 12 | –0.34 | 0.43 | 0.82 | 0.73 | 0.52 |
| 13 | 0.03 | 0.42 | 0.90 | 0.93 | 0.43 |
| 14 | 0.03 | 0.42 | 1.36 | 1.40 | 0.09 |

Table 39. Item Statistics for PC Book 2

| Item # | Difficulty Estimation | Standard Error of Difficulty | Infit | Outfit | Point-Biserial Correlations |
|---|---|---|---|---|---|
| 1 | –3.05 | 0.83 | 1.61 | 1.11 | 0.40 |
| 2 | –2.46 | 0.69 | 1.06 | 1.21 | 0.48 |
| 3 | –3.05 | 0.83 | 0.65 | 0.27 | 0.69 |
| 4 | –4.00 | 1.13 | 0.39 | 0.07 | 0.72 |
| 5 | –3.05 | 0.83 | 0.65 | 0.27 | 0.69 |
| 6 | –2.46 | 0.69 | 0.74 | 0.37 | 0.67 |
| 7 | –0.23 | 0.45 | 0.68 | 0.57 | 0.63 |
| 8 | –0.88 | 0.48 | 1.01 | 0.83 | 0.49 |
| 9 | –0.88 | 0.48 | 0.86 | 0.69 | 0.56 |
| 10 | –1.12 | 0.50 | 1.26 | 1.71 | 0.33 |
| 11 | –0.65 | 0.46 | 1.06 | 0.87 | 0.45 |
| 12 | –0.02 | 0.45 | 0.94 | 0.85 | 0.49 |
| 13 | 0.18 | 0.44 | 1.12 | 1.13 | 0.43 |
| 14 | 0.80 | 0.46 | 1.49 | 1.82 | 0.22 |

Table 40. Item Statistics for PC Book 3

| Item # | Difficulty Estimation | Standard Error of Difficulty | Infit | Outfit | Point-Biserial Correlations |
|---|---|---|---|---|---|
| 1 | –2.55 | 0.62 | 1.53 | 1.22 | 0.18 |
| 2 | –2.20 | 0.56 | 0.75 | 0.61 | 0.60 |
| 3 | –1.90 | 0.52 | 0.58 | 0.49 | 0.72 |
| 4 | –1.90 | 0.52 | 0.60 | 0.70 | 0.69 |
| 5 | –2.20 | 0.56 | 0.57 | 0.35 | 0.72 |
| 6 | –2.20 | 0.56 | 0.57 | 0.35 | 0.72 |
| 7 | –0.99 | 0.43 | 1.20 | 1.41 | 0.29 |
| 8 | –1.19 | 0.45 | 1.69 | 2.19 | –0.02 |
| 9 | –0.99 | 0.43 | 1.03 | 1.05 | 0.40 |
| 10 | –0.31 | 0.39 | 1.04 | 1.04 | 0.36 |
| 11 | –0.99 | 0.43 | 1.27 | 1.27 | 0.25 |
| 12 | 0.16 | 0.39 | 1.06 | 1.60 | 0.24 |
| 13 | –0.15 | 0.39 | 0.81 | 0.68 | 0.52 |
| 14 | 0.01 | 0.39 | 0.90 | 0.78 | 0.46 |

Table 41. Item Statistics for RB Book 1

| Item # | Difficulty Estimation | Standard Error of Difficulty | Infit | Outfit | Point-Biserial Correlations |
|---|---|---|---|---|---|
| 1 | –2.56 | 0.71 | 0.90 | 0.48 | 0.20 |
| 2 | –0.26 | 0.41 | 0.78 | 0.76 | 0.43 |
| 3 | –1.25 | 0.47 | 1.11 | 1.11 | 0.09 |
| 4 | –2.12 | 0.60 | 1.05 | 0.87 | 0.08 |
| 5 | 0.63 | 0.42 | 1.12 | 1.21 | 0.25 |
| 6 | –0.63 | 0.42 | 1.04 | 1.14 | 0.19 |

Table 42. Item Statistics for RB Book 2

| Item # | Difficulty Estimation | Standard Error of Difficulty | Infit | Outfit | Point-Biserial Correlations |
|---|---|---|---|---|---|
| 1 | –3.88 | 1.02 | 0.61 | 0.13 | 0.31 |
| 2 | –1.14 | 0.52 | 1.57 | 1.77 | 0.19 |
| 3 | –0.35 | 0.48 | 1.41 | 1.37 | 0.28 |
| 4 | –1.14 | 0.52 | 0.71 | 0.51 | 0.60 |
| 5 | 0.58 | 0.46 | 0.82 | 0.65 | 0.59 |
| 6 | –0.11 | 0.47 | 0.76 | 0.68 | 0.62 |

Table 43. Item Statistics for RB Book 3

| Item # | Difficulty Estimation | Standard Error of Difficulty | Infit | Outfit | Point-Biserial Correlations |
|---|---|---|---|---|---|
| 1 | –4.57 | 1.75 | 1.00 | 1.00 | 0.00 |
| 2 | –1.38 | 0.54 | 0.84 | 0.55 | 0.52 |
| 3 | –1.09 | 0.51 | 0.98 | 1.26 | 0.43 |
| 4 | –1.38 | 0.54 | 1.26 | 1.10 | 0.34 |
| 5 | 1.32 | 0.50 | 0.74 | 0.56 | 0.62 |
| 6 | 1.06 | 0.49 | 1.24 | 1.18 | 0.43 |

# Appendix 4. Final Text Level Determination Procedure

Instructional reading levels, alternatively known as final text levels, are those text levels at which a student demonstrates instructional oral reading accuracy (90–94%) and comprehension (4 out of 5) or retell/recall performance (2 out of 3). During Atlas field testing, participating students were administered books from consecutive text levels determined to be at or approximate their current instructional reading level as well as multiple books above and/or below their instructional reading level, depending on the specific study design. As a result of this design, students participating in Atlas field testing may or may not reach their final text level as would occur during operational (i.e., nonexperimental) administration of TRC.

The Atlas final text level determination procedure is the same as TRC final text level determination procedure:

- For each book administered, overall book performance is identified as Instructional (INS), Independent (IND), or Frustrational (FRU) based on performance on the TRC components.
- If student performance on the highest administered text level is INS, then the final text level is the highest text level.
- If student performance on the highest text level is FRU, then examine student performance on the next-lower text level (the middle-level text administered in the field study).
- If the middle text level is INS or IND, then the final text level is the middle text level.
- If the highest and the middle text level are both FRU, then examine student performance on the next-lower text level (the lowest text level administered to the student in the field study).
- If the lowest administered text level is INS or IND, then the final text level is the lowest text level administered.

Under usual TRC administration, if the highest text level performance is IND or the lowest text level performance is FRU, test administration continues with additional texts to determine a student's final level. When the administration of books is constrained for research purposes, the result is that there are some cases in which a student's final instructional level could not be determined. Table 33 summarizes this procedure.

Table 44. Final Text Level Determination Rules for the Atlas Field Study

| Lowest Administered Text Level Performance | Middle Administered Text Level Performance | Highest Administered Text Level Performance | Final Text Level |
|---|---|---|---|
| INS/IND | INS/IND | INS | Highest Administered Level |
| INS/IND | INS/IND | FRU | Middle Administered Level |
| INS/IND | FRU | FRU | Lowest Administered Level |
| INS/IND/FRU | INS/IND/FRU | IND | Unable to Determine |
| FRU | INS/IND/FRU | INS/IND/FRU | Unable to Determine |

# For more information visit amplify.com

**Corporate:**
55 Washington Street
Suite 900
Brooklyn, NY 11201-1071
(212) 796-2200

**Sales Inquiries:**
(866) 212-8688 • amplify.com

Amplify.