Identifying Text-Task-Reader Interactions Related to Item and Block Difficulty in the National Assessment for Educational Progress Reading Assessment

Sheila W. Valencia
University of Washington-Seattle

Karen K. Wixson
University of North Carolina at Greensboro

Terry Ackerman
ACT

Elizabeth Sanders
University of Washington-Seattle

October 2017
Commissioned by the NAEP Validity Studies (NVS) Panel

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

James Pellegrino
University of Illinois at Chicago

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington-Seattle

Project Director:
Frances B. Stancavage
American Institutes for Research

Project Officer:
Grady Wilburn
National Center for Education Statistics

For Information:
NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/376-6363

CONTENTS

Reading is interactive and multidimensional (National Institute of Child Health and Human Development, 2000; RAND Reading Study Group, 2002), but reading assessment development is often a linear process with limited attention to the interaction among text, task, and reader factors. As a result, assessment developers may inadvertently misjudge the difficulty of tasks and the nature of the comprehension processes they are seeking to measure. Understanding the full nature of comprehension—the interaction among the text, task, reader, and context—helps us understand how and why comprehension varies from one situation to the next and ultimately allows assessments to capture a more complete measure and interpretation of reading comprehension. As noted in the RAND Reading Study Group, "Until comprehension measures expand to reflect an underlying theory that acknowledges a variety of possible consequences [complex outcomes associated with comprehension–knowledge, application, engagement], both immediate and long term, we will be severely hampered in our capacities to engage in excellent research on this topic" (p. 110).

The purpose of this study was to examine the nature of text-task-reader interactions in relation to student performance on the National Assessment of Educational Progress (NAEP) reading assessment and, in doing so, suggest strategies that NAEP might use to select passages, develop items, and interpret data to more closely approximate the complex nature of reading comprehension.

Background

The 2009 NAEP Reading Framework defines reading for the purposes of assessment. Specifically, it defines reading as "an active and complex process that involves:

- Understanding written text.
- Developing and interpreting meaning.
- Using meaning as appropriate to type of text, purpose, and situation" (p. 2)

NAEP documentation acknowledges that this definition should be considered as a guide for the NAEP reading assessment, not as an inclusive definition of reading. Specifically, "the definition pertains to how NAEP defines reading for the purpose of this assessment" (p. 4).

The 2009 framework also provides ample support for a more complete and complex model of reading by demonstrating that the definition of reading for the NAEP reading assessment is grounded in scientific research on reading summarized in several national reports and related assessments, including the Report of the National Reading Panel (NRP; National Institute of Child Health and Human Development, 2000), *Reading for Understanding: Toward an R&D Program in Reading Comprehension* (RAND Reading Study Group, 2002), the definition of reading that guided development of the Progress in International Reading Literacy Study (PIRLS; Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001), and the definition of reading that guided the development of the Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development,

2000). Equally important is that NAEP's definition is aligned with the Common Core State Standards for English Language Arts (CCSS-ELA) and other college and career readiness standards with similarities to the CCSS-ELA and "what it means to be a literate person in the twenty-first century" (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010, p. 3).

The NAEP framework specifies the types of literary and informational texts that will be sampled at Grades 4, 8, and 12 and also includes attention to genre, subgenres, quantitative measures of readability, and qualitative analyses of text features, such as text structures and literary devices.

Multiple-choice and constructed-response items (tasks) are developed to assess students' comprehension of the literary and informational texts. Each item is further classified by a *cognitive target* that refers to the mental processes or kinds of thinking that underlie reading comprehension. Items assess three cognitive targets: Locate and Recall, Integrate and Interpret, and Critique and Evaluate (National Assessment Governing Board, 2015; see Exhibit 8 on p. 40). The cognitive targets remain the same across all three grades assessed, but the passages and documents on which items are based increase in sophistication at each grade.

Table 1. NAEP Reading Cognitive Targets

	Locate/Recall	Integrate/Interpret	Critique/Evaluate
Both Literary and Informational Text	Identify textually explicit information and make simple inferences within and across texts, such as: • Definitions • Facts • Supporting details	Make complex inferences within and across texts to: • Describe problem and solution or cause and effect. • Compare or connect ideas, problems, or situations. • Determine unstated • assumptions in an argument. • Describe how an author uses literary devices and text features.	Consider text(s) critically to: • Judge author's craft and technique. • Evaluate the author's perspective or point of view within or across texts. • Take different perspectives in relation to a text.

Review of Research

Many studies have addressed individual dimensions of comprehension but, to date, most of the studies focus on text features that can be quantified rather than task and reader dimensions. Equally important, the vast majority of the studies designed to inform reading assessment development examine the parts of an interactive model in isolation from one another rather than simultaneously as they interact in authentic reading situations.

Studies of quantitative indices apply readability formulas (e.g., Fry, Lexile, and Dale-Chall) or other indicators of text complexity (e.g., Graesser, McNamara, & Kulokowich, 2011; Klare, 1984; Stenner & Burdick, 1997) to estimate text difficulty.

Although these formulas are quick and efficient to use, estimates of difficulty can vary substantially when different formulas are used, and efforts to rewrite text to meet readability criteria can lead to incoherence and increased text difficulty (Goldman & Rakestraw, 2000). A recent study of six of the most sophisticated readability schemes explained 36% to 65% of the variance in comprehension (Nelson, Perfetti, Liben, & Liben, 2012). Although correlations of this magnitude suggest that comprehension is shaped by these text indicators, they also suggest that other factors are involved; the most likely candidates are those implicated in an interactive, multidimensional model of reading comprehension.

Other studies have considered reader factors, such as skill/strategy development, language, motivation, and background knowledge (e.g., Alexander, 2007; Johnston, 1984). For example, differences in readers' word reading accuracy and fluency have differential effects on comprehension over the course of reading development (Leach, Scarborough, & Rescorla, 1983; Spear-Swerling, 2004). Similarly, differences in affect and motivation, both intrinsic and extrinsic, influence comprehension as well (e.g., Alexander & Jetton, 2000*;* Braun, Kirsch, & Yamamoto, 2011*;* Guthrie, Wigfield, & VonSecker, 2000), and further interact with topic knowledge and cognitive strategies to produce learning from text (Alexander, 2003).

Finally, others have considered task factors such as type of question, level of thinking, and inference, or specific reading skills associated with particular questions (Bruce, Osborn, & Commeyras, 1994; Davis, 1944, 1968; Valencia & Pearson, 1986). In general, the results of studies of these types of variables are equivocal, with some indicating a strong relationship with comprehension (e.g., background knowledge), others indicating differential effects depending on developing expertise (e.g., interest, motivation, strategies), and still others indicating that performance on specific types of questions or comprehension skills and strategies do not consistently predict comprehension (e.g., NAEP reading stances from the prior NAEP framework, types of questions).

In contrast to the studies above examining single dimensions of reading comprehension, only a few studies and secondary analyses provide emerging evidence of the interaction among various aspects of text, task, and reader, especially as it is understood in reading assessment. In 1997, for example, a special study conducted as part of the 1994 NAEP was designed to compare the performance of Grade 8–12 students who were allowed to select a story to read with those who were assigned to read specific texts (Campbell & Donahue, 1997). The set of comprehension questions for all the stories was identical. A secondary finding from this study, relevant for the work here, found that although none of the stories in either the choice or assigned groups was systematically harder or easier as measured by scores on the questions, students' performance on the identical comprehension questions varied significantly by story. The researchers concluded that there was a text-by-question interaction, suggesting that, "the difficulty of a question resides not only in the question itself, but also in the question's interaction with a particular text" (p. 60).

At about the same time, McNamara, Kintsch, Songer, and Kintsch (1996) investigated the interaction among readers' background knowledge about a topic, the coherence of the text, and the level of understanding required to successfully respond to a test task. Although low-knowledge readers who read high-coherence texts consistently produced

higher comprehension scores than when reading low-coherence texts, the same did not hold true for high-knowledge readers who actually benefitted from minimally coherent texts. Students with high background knowledge who read high-coherent texts produced better results on lower level, recall tasks than on higher level, situational, comprehension tasks. The researchers concluded that low-coherent texts forced high-knowledge readers to compensate and allocate more processing effort, thereby improving their deep understanding, thus highlighting the interaction between knowledge and text.

Approximately 10 years later, Ozuru, Rowe, O'Reilly, and McNamara (2008) investigated the contribution of passage variables (as measured by a subset of Coh-Metrix variables) and question variables (representing depth of processing) on middle and high school reading test items, asking "Where's the difficulty in standardized reading tests: the passage or the question?" They found that item difficulty at middle school was primarily influenced by text features rather than question type; researchers identified several interactions between text and question variables, but these were not interpreted or analyzed with respect to interactive theories of comprehension. In contrast to the findings for middle school students, neither text nor question variables, as measured in this study, accounted for a significant percentage of the variance in comprehension scores at the high school level. In other words, the variables used in this study to predict the comprehension difficulty of 12th-grade students had no relation to actual text difficulty and student comprehension. These findings suggest that predictors of comprehension for high school-aged students were not picked up by the variables used in this particular study.

A more recent study by Ogut, Dogan, Tirre, Ndege, and Hummel (2010) investigated text and question factors that impacted the comprehension difficulty of fourth- to eighth-grade NAEP passages and distinguished performance among particular subgroups of students. Researchers identified three categories of difficulty: (1) a subset of Coh-Metrix passage factors (different than those used in Ozuru et al., 2008), (2) a simple item-type variable (open-ended or multiple-choice), and (3) a simple genre variable label. Controlling for item type and genre, they found that some of the passage variables (i.e., Coh-Metrix factors) were significant predictors of overall fourth-grade reading difficulty and score differences between subgroups. However, none of these factors was significant at Grade 8. It is important to note that, similar to the Ozuru et al. (2008) study, this study did not address interactions among text, task, and reader variables in their factors or analyses. Furthermore, as in the Ozuru et al. study, there was a differential contribution of factors across grade levels that was not explored. Across both these studies, we are left wondering which factors actually account for reading performance, how they interact as readers engage with text, and how they influence comprehension at different grade levels.

Two studies, based on adult literacy assessments, are most closely aligned with an interactive, multidimensional approach to developing and interpreting reading assessments. Kirsch (2001) describes how an interactive model was used to conceptualize and construct the International Adult Literacy Survey (IALS). Although this assessment includes real-world adult-level prose, document, and quantitative literacy tasks that pose different demands than are typical for younger students, Kirsch's approach models how an interactive perspective might be employed in assessment design.

Specifically, Kirsch (2001) describes three categories of factors that were manipulated in the development of IALS tasks—adult contexts/content, texts, and process/strategies—and a number of variables within each. Three process/strategy variables were identified (i.e., type of match, type of information requested, and plausibility of distractors), all of which consider the item (task) demands with respect to the specific text being read. A regression analysis on prose comprehension, including these three variables and a traditional readability score, found that only type of match (i.e., what was asked in the questions and what/how it was presented in the text) was significant. Nevertheless, Kirsch pointed out that although each of these variables may not be significant in terms of regression analysis, "each was taken into consideration when constructing the literacy tasks and, therefore, each is important as to how well the domain is represented" (p. 32).

Following from Kirsch's (2001) work, White (2011) conducted a secondary analysis of data from the 2003 National Assessment of Adult Literacy (NAAL) and the 1992 National Adult Literacy Survey (NALS). Using multidimensional item response theory (IRT), White proposes a theory to explain functional adult literacy in terms of the interaction of text features, task demands, and related reader skills. Although she acknowledges that aspects of a particular text and task may independently affect the ease or difficulty of comprehension (e.g., the length of a text, mathematical operations to be performed), she argues that the interface between the text and the task is central to consider—"both text and task features influence task demands" (p. 69). Furthermore, she argues that it is possible to analyze the task demands and their relation to the text independent of a reader's abilities. Finally, White has identified what she calls task facilitators and inhibitors—task-relevant features or portions of a text that assist or divert readers from successfully accomplishing a specific reading task. She provides a helpful example:

> Many variables can either aid or hinder performance, depending on the specifics of the particular task and text. For example, suppose that the year 1978 is italicized in a text associated with a literacy task. If 1978 is the correct answer the italics might make it easier for the reader to find the answer. However, if the correct answer is 1979, the italics might draw respondents' attention to 1978, the wrong year. Text features interfacing with the task are facilitators when they aid task performance and inhibitors when they hinder task performance (p. 71).

Finally, extending the work of Kirsch and White to younger students, and as a precursor to the current study, Valencia, Wixson, and Pearson (2014) examined the relation among text, task, and reader using extant NAEP-released data. Their aim was simply to lay the groundwork for further study into these relations as they play out on reading assessments. They demonstrated the interactive nature of text, task, and reading by highlighting differential item difficulty on items (tasks) related to a single text (with a single readability index), and the ability to manipulate texts and tasks within a grade-level reading block to both increase or decrease comprehension difficulty.

As this overview of research reveals, much of the work on comprehension and comprehension assessment has focused on single facets of an interactive model of reading. Studies of text factors typically focus only on the text itself, omitting attention

to reader and task variables, including the various types of questions asked about a text. Similarly, studies of readers' abilities to answer a variety of questions rarely examine text factors systematically to determine how they interact with task demands. Efforts to systematically examine the interactive nature of reading performance on reading assessments are just emerging. Further studies are needed to provide more theoretically driven, complex models of assessment development and interpretation.

The Study

This study addresses the construct and instructional validity of the NAEP reading assessment. It aims to sharpen the lens used to identify passages and develop items so that it is more consistent with the lens used in current curriculum design efforts, educators' judgments about texts and tasks, and new large-scale assessment development efforts, including those intended to evaluate achievement of college and career readiness standards in English language arts.

Building on prior research, we consider three interrelated areas: (1) quantitative and qualitative dimensions and characteristics of *text*, (2) qualitative and quantitative aspects of *tasks* (i.e., items and blocks), and (3) actual *reader* performance (Valencia, Pearson, & Wixson, 2011). By examining these simultaneously, we aim to identify features that could assist NAEP personnel in selecting passages, constructing items, and interpreting findings in line with a multidimensional, interactive model of reading supported by research. In addition, the study may contribute to a more nuanced understanding of how instruction could address students' abilities to comprehend what they read in relation to a variety of text-task-reader combinations.

Specifically, we address the following questions:

a. Which qualitative and quantitative characteristics designed to examine text-task-reader interactions are related to comprehension difficulty at Grades 4, 8, and 12?
b. What differences among these characteristics, if any, exist across blocks, grade levels, genres, item types, and/or cognitive targets?

Method

Data Set

The data set consisted of eight blocks from the 2009 NAEP reading administration. Each of these blocks included one reading selection and associated multiple-choice and constructed-response items. Blocks with more than one reading selection were not included because those available to us generally used poetry as the second text and none of the quantitative readability formulas are designed to be applied to poetry. Five of the blocks used in this study were administered across two grade levels and thus provided for an examination of developmental differences. In all, data for a total of 13 blocks were analyzed. Table 2 shows the block, genre, and number of items associated with each. Items that targeted specific text vocabulary were not included in this study.

Table 2. Descriptive Variables of NAEP Reading Blocks Used for Analysis

Block Number	Block "Name"	Grade	Genre**	Number of MC Items	Number of CR Items
1	LW*	4	I	4	4
2	WH*	4	I	5	3
3	DA	4	L	4	4
4	LN*	4	L	5	3
5	LW*	8	I	4	4
6	WH*	8	I	5	3
7	LN*	8	L	5	3
8	DR*	8	I	3	5
9	DU*	8	L	3	5
10	SC	12	I	5	4
11	OW	12	L	3	6
12	DR*	12	I	3	5
13	DU*	12	L	3	5

*Cross-grade blocks
**I=informational, L=literary
Note: MC=multiple choice, CR=constructed response

For each block, we were provided with the reading passage, Lexile and TextEvaluator readability scores, and individual items with scoring keys/rubrics and cognitive target classifications. In addition, we had access to NAEP student performance data from 2009, including the percentage of students selecting each option for multiple-choice items, the percentage of students scoring at each score level for constructed-response items, and IRT difficulty parameters.

Procedure

We developed two types of text-task-reader characteristics using existing research findings to identify characteristics of texts and tasks that had been shown to be associated with comprehension performance. The first type, Text-Task-Reader Comprehension (TTR-C) characteristics, focuses on the interaction among the text, item stem, and correct response. The second type, Text-Task-Reader Distractor characteristics (TTR-D), focuses on the characteristics of the distractors used in multiple-choice items in relation to the text, item stem, and correct response. The following describes the final set of characteristics we developed after a series of iterative analyses. These are the characteristics used in the analyses in this study.

TTR-C Characteristics. After a careful review and content analysis of each reading block (reading passage, items, and answer keys/rubrics), reading passages and items were examined further in relation to IRT difficulty parameters and overall percentages of correct responses for each item, as well as the percentage of respondents selecting each distractor for multiple-choice items or the percentage scoring at each level for constructed-response items. The goal here was to work

backward from the data to develop an initial scoring rubric that would represent the interaction among text, item (task), and reader.

Working iteratively from the existing database, we developed a list of TTR-C characteristics and a scoring rubric to evaluate items. The characteristics were designed to consider reader, text, and task (item) influences simultaneously, reflecting what readers must read, understand, and do to respond correctly to a specific task (reading assessment item). A draft scoring rubric was presented at the May 2014 National Validity Studies (NVS) Panel meeting and then discussed at the NVS meeting in September 2014 after it had been applied to several blocks and revised accordingly. Based on feedback at the September meeting, a panel of subject-matter experts was identified to assist the two study lead authors in refining the rubric starting in the fall of 2014 (see Appendix A for the list of experts). The group convened three times over 3 months to revise the characteristics, definitions, and rubric. Characteristics were deleted, added, and refined, and rubrics were rescaled, resulting in improved validity and inter-rater reliability for a subset of "training" blocks—two cross-grade blocks (an informational Grade 4/8 block and a literary Grade 8/12 block). The final list of TTR-C characteristics is shown in Table 3 and the final scoring rubric is included in Appendix B.

Table 3. TTR-C Characteristics

A. Reading processes related to text-task-reader interactions

1. *Cognitive Complexity*—Degree of cognitive complexity to go from the stem to text-based inference(s) and back to the correct answer.
2. *Abstractness*—Abstractness of text content in relation to the key/rubric or process of abstract reasoning to get to the key.
3. *Amount of Text/Source of Information*—Minimum amount of text/source of inference needed for the correct answer; no prereading assumed.
4. *Synthesis*—Process of pulling together information or understanding in the text or to go from the text to the correct answer.

B. Item characteristics related to text-task-reader interactions

1. *Stem/Key/Text Language Alignment*—Extent to which there is shared or common language among the stem, key, and text that leads to a correct response.
2. *Stem Directness/Cueing*—How much specificity/cueing the item stem provides to the location or identification of the correct answer.
3. *Response Elaboration*—Amount or number of pieces of information from the text required by the item to get full credit.

C. Text characteristics related to text-task-reader interactions

1. *Text Structure/Genre/Organizational Pattern*—Extent to which the structure of the text (genre, organization) is a factor in arriving at the correct answer.
2. *Text Format/Features*—Extent to which the format of the text (i.e., text features such as headings, illustrations, and boldface) is a factor in arriving at the correct answer.
3. *Text Language*—Extent to which text language (e.g., concreteness/abstract concepts, vocabulary, syntax, colloquial language, dialect) is a factor in arriving at the correct answer.
4. *Literary and Rhetorical Style*—Extent to which literary and rhetorical features (e.g., imagery, metaphor, characterization, foreshadowing, persuasive language, literary terms, theme) are factors in arriving at the correct answer.

Next, the items for all the blocks in the study were scored, including rescoring the initial "training" blocks, using the revised scoring rubric. At two face-to-face meetings, four members of the original expert team were oriented and trained to use the new scoring procedure and rubric. Thereafter, each block was systematically assigned to be scored by three of the four raters. Intraclass correlations (ICCs) revealed an acceptable level of reliability in coding across all blocks used in this study (see Table 4).

TTR-D Characteristics. As others have demonstrated, distractors play a major role in the difficulty of multiple-choice items (Kirsch, 2001; White, 2011). Building on text-task-reader interactions, we developed a set of distractor characteristics and then qualitatively analyzed "strong" distractors—those that drew 14% or more of the respondents. As with the TTR-C characteristics, this analysis considered each "strong" distractor in relation to the text, the item, and the correct keyed response.

Distractor characteristics were developed iteratively by analyzing every "strong" distractor identified across a total of 104 items in the 13 blocks. Each distractor was assigned as many characteristic codes as applied. Two raters independently assigned codes; differences were adjudicated through discussion and refinement of distractor characteristic definitions. Table 5 describes the distractor characteristics associated with multiple-choice items.

Table 4. Intraclass Correlations (ICCs) for TTR-C Characteristics†

Characteristic	ICC
Cognitive Complexity	.96***
Abstractness	.92***
Amount of Text	.98***
Synthesis	.94***
Stem/Key/Text Alignment	.95***
Stem Directness/Cueing	.97***
Response Elaboration	.99***
Text Structure	.95***
Text Format/Features	.98***
Text Language	.77***
Literary/Rhetorical Style	.96***

†Two-way mixed-effects model where people effects are random and measures effects are fixed. Type A ICCs using an absolute agreement definition.

*** $p < .001$

Table 5. Qualitative Multiple-Choice TTR-D Characteristics

a. Distractor appears in the first position (choice a).
b. Distractor is logical or reasonable but incorrect (could be deduced simply by logic, with or without information in the passage).
c. Distractor choice reflects misunderstanding of linguistic or literary issues in the text or distractor (e.g., vocabulary, syntax [anaphora], figures of speech, literary style, structure [flashback], perspective).
d. Distractor includes words that are identical to key words in the text.
e. Distractor includes information about an actual event or idea presented in the text.
f. Information in the distractor appears multiple times in the text.
g. Distractor is partially correct based on information in the passage but does not represent a complete response or understanding.
h. Distractor information is close to stem words in the text, but the correct response is farther away from stem words in the text.
i. Distractor represents a literal response to a conceptual question (e.g.. main idea, why is X important to the story, main way the author shows, author evidence to support points, main strategy the author uses). Correct response requires a generalization.
j. Distractor presents an "obvious" choice, whereas the correct response requires an inference that may be missed or is not as obvious.[1]

Analyses

Data analyses began with descriptive and exploratory analyses to establish the viability and validity of the TTR-C characteristics. Second, we conducted quantitative analyses at the block and item levels using a two-level hierarchical linear model (HLM) to predict the dependent variable-percent correct for each item. Third, we conducted fine-grained qualitative analysis for blocks at each grade level to examine the variability of TTR-C characteristics in an effort to describe patterns within and across grades. We also qualitatively analyzed the variability of the TTR-D characteristics for the "strong" distractors for every multiple-choice item—those that drew more than 14% of respondents—to gain another perspective on students' thinking and comprehension of text. Finally, we created and examined profiles of different blocks to more closely examine the qualitative variability of both the TTR-C and TTR-D characteristics within and across items and blocks. Because the IRT b parameters for item difficulty are calculated within grade level, p+ statistics (the percentage of students scoring correct responses) were used in all cross-grade analyses. The b parameters were used for within-grade analyses.

Comparison of Text/Block Difficulty

To confirm the basic premise that text variables included in readability indices do not adequately capture the difficulty that results from text-task-reader interactions,

[1] Considered but not scored (too few distractors aligned with these characteristics):
- Question is out of sequence with respect to other questions and the text.
- Question asks for a conceptual response, but the literal answer is keyed as correct (and the conceptual distractor is keyed as incorrect).

comparisons of difficulty were created for the blocks that were administered at each grade. Using standardized scores, each block, labeled by "name" as in Table 2, was arrayed according to difficulty as measured by two text readability measures (Lexile and TextEvaluator) and by the average difficulty parameter of the items in each block. Like most other quantitative measures, Lexile and TextEvaluator methods use algorithms to produce scores that indicate a progression of difficulty according to specific text features, such as sentence length, word frequency, word types, word syntactic features, and so on, although the two readability measures used here apply different procedures and factors in their calculations (see Nelson, Perfetti, Liben, & Liben, 2012). Figures 1, 2, and 3 display the comparisons by grade level for ease of interpretability.

Overall, the comparisons reveal inconsistent patterns of placement of the blocks with respect to "difficulty" across actual student performance (average item difficulty) and each text readability score. Although the two readability formulas were significantly correlated with each other ($p<.05$), average item difficulty for each block was not significantly correlated with either readability score for the associated reading selection. These analyses suggest that factors other than those included in readability measures are contributing to comprehension performance and difficulty. These "other" factors, drawn from research, are the basis of our TTR-C and TTR-D characteristics. Furthermore, in line with our hypothesis about the interaction of text, task, and reader, we expect that the misalignment of difficulty indicators that we see in Figures 1, 2, and 3 would vary with different test items for these reading passages or different samples of students responding to them.

Figure 1. Comparison of Difficulty Measures for Grade 4 Blocks

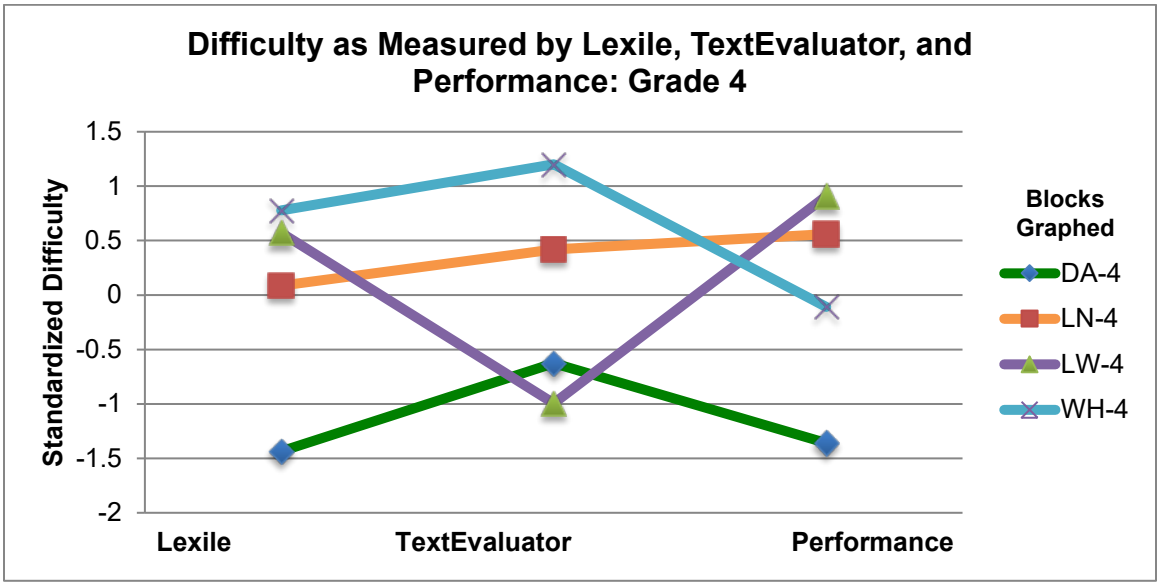


Figure 2. Comparison of Difficulty Measures for Grade 8 Blocks

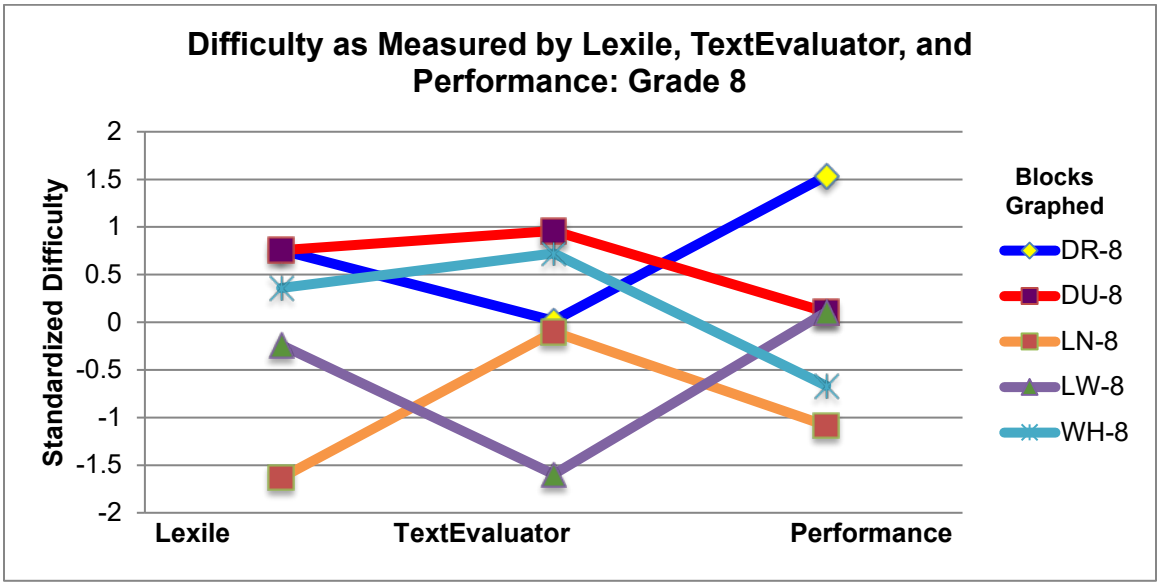
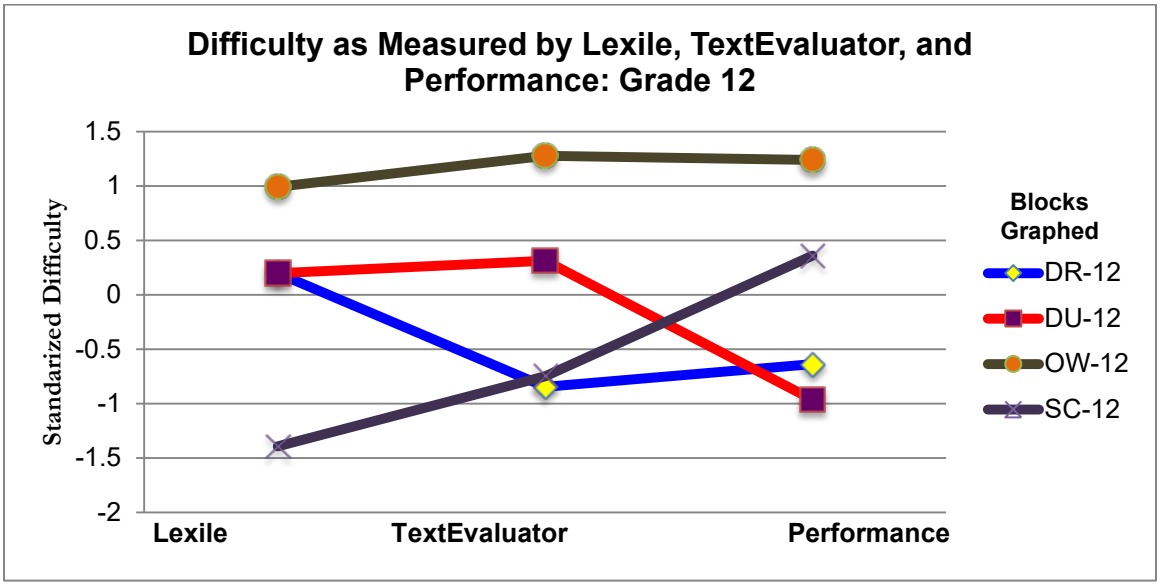


Figure 3. Comparison of Difficulty Measures for Grade 12 Blocks



Examining Block, Item, and TTR-C Characteristics

Quantitative analyses at the block and item levels were conducted using a two-level HLM (also known as linear mixed models) on the dependent variable—*percent correct* for *each item;* that is, the percentage of students scoring correct responses (PC, range: 0.00–100)—to indicate item difficulty, with items at Level 1 (104 items) and blocks at Level 2 (13 blocks). This analysis was designed to (a) account for the nonindependence of items within blocks as well as (b) test the item- and block-level predictors with correct degrees of freedom. Note that *percent score* (PS, range: 25–100) was used for item-level *characteristic predictors* (rather than raw points) to maintain a similar scale, as some of the predictors varied in the maximum number of points possible (see the rubric in Appendix B). In addition, for further brevity, all

categorical predictors were effect-coded (so that the coefficients represent the effect of the predictor compared with the mean and the value of the coefficient should be doubled to determine differences between categories) and all metrical predictors were standardized (in z-scores). Finally, we note that all models were estimated using restricted maximum likelihood in *HLM7*.

The PS means and standard deviations for each characteristic are presented by grade level in Table 6. Overall, scores for most of the characteristics were reasonably distributed across the range of possible scores on the rubric; the exceptions were text format/features and text structure. The limited range of scores in these categories reflects both reading passages and questions in the sample of blocks used for this study—text features and text structure were not factors in understanding the text or answering the questions. Furthermore, the features simply were not present in the texts in many of the blocks. Nevertheless, these characteristics were important considerations when a block *did* contain texts and/or items that included or required understanding of these features. This is explored further in the qualitative analysis.

Table 6. Means (PS)[1] and Standard Deviations for TTR-C Characteristic Scores by Grade Level

Characteristic	Grade 4 *N*=32		Grade 8 *N*=39		Grade 12 *N*=33	
	X	SD	X	SD	X	SD
Cognitive Complexity (Cog complx)	64.1	24.5	64.7	26.1	68.2	22.8
Abstractness (Abst)	64.6	25.5	63.3	26.4	72.8	25.7
Amount of Text (Amt)	73.9	32.7	74.3	32.1	77.7	29.3
Synthesis (Synth)	63.5	29.9	62.4	27.9	66.7	27.8
Stem/Key/Text Alignment (Align)	60.2	25.3	62.8	27.4	70.5	26.8
Stem Directness/Cueing (Direct)	70.8	29.2	70.9	28.9	70.7	29.9
Response Elaboration (Elab)	50.9	25.6	49.4	24.2	47.3	22.3
Text Structure (TxtStruc)	39.3	15.9	40.8	16.4	36.1	9.9
Text Format/ Features (TxtFeat)	39.3	15.9	38.2	14.6	33.0	00.0
Text Language (TxtLang)	35.1	8.4	36.5	10.5	41.2	14.8
Literary/Rhetorical Style (Lite/Rhet)	36.2	10.1	38.2	14.6	54.4	28.8

[1]Possible PS range for all scores: 10.00–100
Note: X=mean, SD=standard deviation

Item Characteristic Principal Component Analysis. To potentially reduce the number of item characteristics into composites prior to modeling characteristic links with item percent correct, a principal component analysis (PCA) was conducted on the item characteristics. The PCA found four components (linear combinations of the variables) with eigenvalues greater than 1. An orthogonal rotation was employed for ease of interpretation of the loadings. With a critical value of $r = \pm 0.51$ for a sample size close to $N = 100$ (Stevens, 2002, p. 394), only one composite (comprising Cognitive Complexity, Stem/Key Text Alignment, Abstractness, Synthesis, Amount of Text, Stem Directness/Cueing, and Response Elaboration)

was found to be reliable. The other four characteristics that did not load on the composite were treated as separate variables.

Indeed, the seven TTR-C characteristics found to correlate with this component have an internal consistency reliability estimate (Cronbach's alpha) of 0.899. The other four TTR-C characteristics (Text Structure, Text Format/Features, Text Language, and Literary/Rhetorical Style) had no substantive or significant correlations with each other and were in fact the only characteristics that were found to have a non-zero ICC (see next section); in other words, these characteristics that were not part of the component exhibited variability by block. Thus, the only composite created was for the first seven characteristics (by taking the mean across those variables).

Block Variation in the PC Outcome and Item Characteristic Predictor Variables. For each variable of interest, we conducted separate two-level HLMs with no predictors (called an "unconditional" or "intercept-only" model) to test whether there was significant variance between blocks on the outcome, item PC, and each of the item characteristic predictors. The results of the fixed-effect portion of these models (see Table 7, *Intercept* columns) showed that the mean of the item PC outcome, controlling for block variation, was 62% (item PC *Coeff* = 0.62, $p < 0.001$), and that the predictor variable means, controlling for block variation, ranged from a low of 37% (TxtFeat *Coeff* = 0.37, $p < 0.001$) to a high of 75% (Amt *Coeff* = 0.75, $p < 0.001$). (In fact, all of the means were significantly greater than zero.) More interestingly, the random-effect portion of these results (see Table 7, *Variance* columns) show that there was significant variance between blocks in item PC as well as for three of the item characteristics (TxtStruc, TxtFeat, and Lit/Rhet), all p-values < 0.05. In particular, the estimated ICCs (variance between blocks divided by total variance) for item PC, TxtStruc, TxtFeat, and Lit/Rhet were 0.27, 0.11, 0.28, and 0.33, respectively. These values indicate that block accounted for 27% of the variation in item PC, and 11%, 28%, and 33% of the variance in the respective item characteristics. Although not significant, TxtLang also exhibited nonzero block variation, with 3% of the variability in this item characteristic accounted for by block. These characteristics are more heavily influenced by the nature of the text included in the block than the other TTR-C characteristics. For example, if there are no text features associated with the reading selection or the text structure is not unusual in any way (i.e., featuring flashbacks), then questions will not address these characteristics and they will not be reflected in text-task-reader interactions. No block variation was detected on the first seven characteristics or the composite.

Subsequent analyses were performed using both individual item and block characteristics and sets of item and block characteristics. The item variables consisted of the TTR-C characteristics, item type, and cognitive target. The block variables consisted of quantitative readability measures (Lexile and TextEvaluator), grade level, and genre.

Table 7. Intercept-Only Model Results to Evaluate Variation Between Blocks

Variable	Intercept (Mean)				Variance				
	Coeff	*SE*	*t(12)*	*p*	*Between Blocks*	*Within Blocks*	$\chi^2(12)$	*p*	*ICC*
Outcome									
Item PC	.62	.03	23.10	< .001	.0071	.0194	46.95	< .001	.27
Characteristic Predictors									
Cog complx	.66	.02	27.39	< .001	.00	.06	1.83	> .500	.00
Abst	.67	.03	26.15	< .001	.00	.07	10.66	> .500	.00
Amt	.75	.03	24.42	< .001	.00	.10	10.45	> .500	.00
Synt	.64	.03	23.09	< .001	.00	.08	9.44	> .500	.00
Align	.64	.03	24.61	< .001	.00	.07	7.64	> .500	.00
Direct	.71	.03	24.88	< .001	.00	.08	6.92	> .500	.00
Elab	.49	.02	21.02	< .001	.00	.06	2.75	> .500	.00
Composite (first 7 char.)	.65	.02	30.98	< .001	.00	.05	6.34	> .500	.00
TxtStruc	.39	.03	20.46	< .001	.00	.02	24.15	.019	.11
TxtFeat	.37	.02	17.05	< .001	.00	.01	49.26	< .001	.28
TxtLang	.38	.01	29.68	< .001	.00	.01	14.59	.264	.03
Lit/Rhet	.42	.04	11.38	< .001	.01	.03	63.19	< .001	.33

Note: SE=standard error; ICC=intraclass correlation

Links Between Item- and Block-Level Predictors and Item PC Outcome. Four separate analyses were conducted to evaluate the relationships among block- and item-level predictors and the item PC outcome (see Table 8 for results). First, a series of models with each of the block- and item-level predictors entered individually was conducted to evaluate the "direct" effect of each predictor on the outcome (item PC), similar to a zero-order correlation but also accounting for blocks. (For this first set of models, no intercept is shown as the intercept estimate varies slightly depending on the predictor entered into the model; however, we note that it is always approximately 62%.) Next, a series of models was employed to test the unique contributions of the *set* of block characteristics (Model 1)[2] and the *set* of item characteristics (Model 2).[3] Finally, in Model 3, we estimate the unique contributions of the item-level characteristics, after controlling for block-level predictors. Note that, in these analyses, all categorical predictors were effect coded and all continuous predictors were standardized into z-scores for ease of results interpretation. For example, Item Type is effect coded as -1=multiple choice and 1=constructed response; Grades 8 and 12 are an effect-coded set of two predictors, with Grade 4 as a reference group (coded -1); Cog Target Locate/Recall and Cog Target Critique/Evaluate also are an effect-coded set, with Cog Target Integrate/Interpret as a reference group. Coefficients for these effect-coded variables are interpreted as the effect of that item type, grade level, or cognitive type as different from the mean percent correct for items (PC), holding all else constant. For brevity, the table of

[2] Block characteristics used in Model 1 are Grade Level, Passage Type, Lexile, and TextEvaluator.
[3] Item characteristics used in Model 2 are Item Type, Cog Target, Composite 1, TxtStruc, TxtFeat, TxtLang, and Lit/Rhet.

model results (Table 8) displays fixed effects. (Between-block variation was accounted for as a random effect in each model; random effects available from authors.)

Table 8. Full Model Results Comparing Direct and Unique Effects

			Indirect Effects (Unique)					
	Direct Effects (Individual)		Model 1 (Block)		Model 2 (Item)		Model 3 (Block+Item)	
Fixed Effects	*b*	*P*	*b*	*P*	*b*	*p*	*b*	*p*
Intercept (Mean)	—		.62	< .001	.61	< .001	.61	< .001
Block-Level Predictors								
Grade 8	.06	.059	.07	.033			.06	.020
Grade 12	.04	.196	.05	.161			.08	.009
Psg Type	-.01	.850	-.03	.230			-.01	.581
Lexile (z)	.01	.637	-.07	.082			-.05	.075
Text Eval (z)	.03	.352	.04	.224			.05	.083
Item-Level Predictors								
Item Type	-.08	< .001			-.06	< .001	-.06	< .001
CogT L/R	.04	.056			-.07	.014	-.06	.015
CogT C/E	-.06	.003			.04	.146	.03	.167
Cog complx(z)	-.05	< .001						
Abst (z)	-.05	< .001						
Amt (z)	-.03	.021						
Synt (z)	-.05	.001						
Align (z)	-.05	.001						
Direct (z)	-.02	.232						
Elab (z)	-.08	< .001						
Composite (z)	-.06	< .001			-.05	.011	-.05	.015
TxtStruc(z)	-.02	.228			-.01	.493	-.01	.536
TxtFeat (z)	-.03	.031			.00	.715	.00	.831
TxtLang (z)	-.03	.052			-.04	.003	-.04	< .001
Lit/Rhet (z)	-.06	< .001			-.04	.016	-.05	.002

Note: Fixed effects shown; random effects available upon request.

Block Variable Effects. Results of the direct-effects model indicated that none of the block-level predictors had significant relationships with the outcome variable (item PC) when entered individually. (Recall that the intercept is not shown as the conditional mean of the outcome, item PC; the intercept varied for each of these models depending on which predictor was entered, but was generally a mean of 62%.) However, when entered as a set of predictors together (see Unique Effects Model 1), Grade 8 had a significant positive relationship with the outcome (i.e., item percent correct—PC): in other words, Grade 8 items had significantly more students scoring correct—7% higher than the average; Grade 12 items trended in the same direction, although to a lesser extent. This indicates that, overall, Grade 4 blocks were relatively more difficult than blocks at either Grade 8 or Grade 12. Based on the coefficients from the direct model, predicted values for Grade 4 items are estimated to be 10% below average percent correct.

These results seem reasonable considering that three of the four blocks analyzed at Grade 4 also were administered at Grade 8 and may be more challenging for Grade 4 than others; similarly, two of the four blocks analyzed at Grade 12 were administered at Grade 8. The results also indicate that neither of the block-level readability estimates (Lexile, TextEvaluator) was significantly related to item difficulty. This finding mirrors the descriptive comparison graphs (Figures 1, 2, and 3), which show inconsistent measures of difficulty across readability estimates and actual item difficulty results. Furthermore, this analysis found no significant difference by Passage Type (genre); percent correct was not significantly different across literary and informational passages. Using the intercept-only model total variance estimate compared with the total variance estimates for Model 1, block characteristics were found to account for approximately 18% of the variance in percent correct (as a set).[4]

Item Variable Effects. The direct-effects model results also show that most item-level predictors had significant negative relationships with item percent correct (PC) when entered individually (i.e., the more challenging the characteristic, the lower the PC and the greater the item difficulty). Exceptions were Cognitive Target L/R and three of the TTR-C characteristics (Direct, TxtStruc, and TxtLang). When item-level predictors were entered together as a complete set using the composite, the remaining four TTR-C characteristics (TxtStruc, TxtFeat, TxtLang, and Lit/Rhet), item type, and cognitive target all were significant except for CogT C/E, TxtStruc, and TxtFeat (see Unique Effects Model 2). Using the intercept-only model total variance estimate compared with the total variance estimates for Model 2, item characteristics, as a set, were found to account for 25% of the variance in percent correct.

To interpret the reversal in direction of the two cognitive target predictors when included with all the other item-level predictors, we looked further. These predictors turned out to be highly correlated with each other (due to their common reference group and unequal category sizes, at $r = .63$), and also are correlated with the composite, at $r = -0.41$ and $r = 0.26$ for CogT L/R and CogT C/E, respectively. CogT C/E also is positively correlated with TxtStruc, TxtFeat, and Lit/Rhet, at $r =$ 0.30, 0.25, and 0.20, respectively. These complex intercorrelations are a likely reason the relationship signs for CogT L/R and CogT C/E shift in opposite directions when the set of item characteristics are entered together. Further analysis using a one-way analysis of covariance (ANOVA) with Tukey's pairwise comparisons showed that there was a main effect of cognitive target on percent correct (F-test p-value = 0.029), that CogT L/R and CogT I/I (the reference group) were not significantly different from each other (q-test adjusted p-value = 0.944), and that those two groups appeared to differ from CogT C/E (L/R versus C/E-adjusted $p =$ 0.057; I/I versus C/E-adjusted $p = 0.047$). Overall, the cognitive target codes are heavily related to the other item characteristics, and the sign change may be an artifact of multicollinearity with the composite variable.

[4] Total variance for the intercept-only model was used to calculate the contributions of the block-level and item-level predictors separately and jointly (total variance in percent correct using intercept-only models is .0071 between blocks, + .0194 within blocks = .0265).

Additional analyses were conducted to examine TTR-C characteristics across the three levels of cognitive targets. Multivariate analysis revealed significant differences in mean scores for the composite (p-value = 0.000), across all pairwise comparisons (i.e., CogT L/R versus CogT I/I; CogT L/R versus CogT C/E, CogT I/I versus CogT C/E). Significant pairwise differences also were found for TxtStruc (CogT Locate/Recall versus Critique/Evaluate, p-value = 0.043; Integrate/Infer versus CogT C/E, p-value = 0.030) and for Lit/Rhet (CogT L/R versus CogT C/E, p-value = 0.005). These findings suggest that TTR-C characteristics may be useful in refining definitions and developing items aligned with particular cognitive targets.

We also looked further to understand the relationship of TTR-C characteristics with item type. Following on the finding that constructed-response items were answered correctly (i.e., full-credit responses) by significantly fewer students than multiple-choice items, we found that the composite, TxtFeat, and Lit/Rhet were significantly correlated with percent correct for constructed-response items; only TxtLang was significantly correlated with percent correct for multiple-choice items. There were significant differences in the mean TTR-C characteristic scores of multiple-choice and constructed-response items for the following characteristics: Cog complx, Abst, Align, Elab, TxtFeat, and Lit/Rhet. Scores for constructed responses on these characteristics were higher than for multiple-choice, indicating more challenge on these TTR-C characteristics.

Finally, full model results with both block and item characteristics (Model 3) were estimated. Those results were consistent with the prior models for block and item variables as sets in isolation. Using the intercept-only model total variance estimate compared with the total variance estimates for Model 3, the block and item variables jointly accounted for approximately 59% of the variance in percent correct, which accounts for 41% more than the block variables alone.[5] In interpreting this finding, we caution about two issues with regard to our models: (1) variance estimates from maximum likelihood estimates, such as those used by our *HLM* software, are prone to underestimation and therefore effect sizes are approximate; and (2) our outcome (percent correct) variance is near a zero-boundary as it is a percentage value. Given this, the estimated "variances accounted for" are likely to be overestimated.

Overall, the results of the HLM analysis, as well as theoretical and empirical models of reading comprehension, suggest differences in the relationship between many TTR-C item characteristics and item percent correct (item difficulty) across grades, item types, cognitive targets, and individual blocks. In addition, visual inspection of scatterplots across blocks indicates considerable variability in the characteristics that are most strongly related to item difficulty within each block—they vary as a function of the text-task-reader interactions. The nonsignificant results of readability measures and text type (genre) on item difficulty further support the possibility that TTR-C characteristics might account for some of the variance in item difficulty. Therefore, to more thoroughly understand TTR-C characteristic effects, additional analyses were pursued within each grade level as described below.

[5] The total variance of Model 3 summed to .0108 compared with the intercept-only model's total variance of .0265.

Grade-Level Analysis

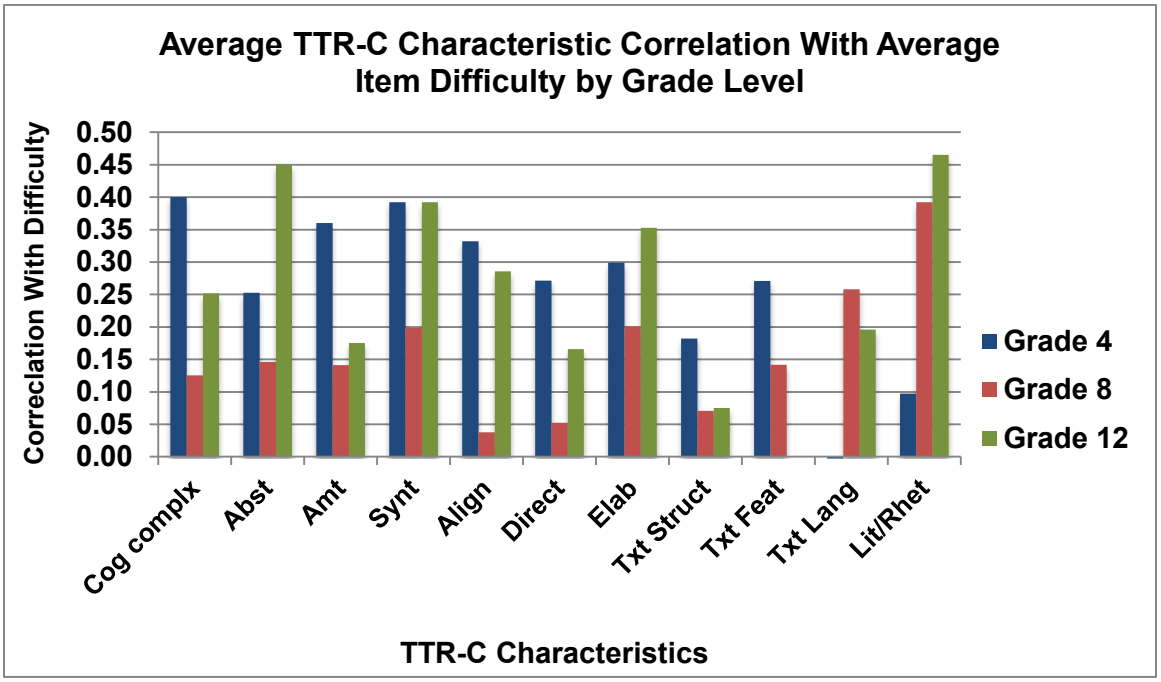
Text-Task-Reader Comprehension (TTR-C) Characteristics

To explore how particular characteristics were related to item difficulty at each grade level, we conducted exploratory quantitative and qualitative analyses. Correlations between mean characteristic scores and mean item difficulty as measured by IRT difficulty parameters revealed that the following characteristics were significantly related to difficulty at each of the grade levels.

- Grade 4—Cognitive Complexity, Amount of Text, Synthesis
- Grade 8—Literary/Rhetorical Style
- Grade 12—Abstractness, Synthesis, Response Elaboration, Literary/Rhetorical Style

These results, along with an examination of scatterplots and correlation graphs of the TTR-C characteristics within grade level (see Figure 4), revealed a great deal of variability within and across grades, prompting further exploration of the variability of the characteristic scores within blocks. It is important to interpret these data in terms of text-task-reader interactions rather than simply as individual characteristics; they indicate the characteristics that are required of readers who correctly answer specific items (tasks) for specific texts within a block. Because blocks vary substantially in the nature of texts and associated tasks (items), these characteristics may not be equally represented in each block. For example, at fourth grade, there are fewer items that require readers to understand authors' use of literary/rhetorical features than at 12th grade, thus potentially impacting the correlation. Therefore, these patterns cannot be interpreted to suggest developmental reader differences or indicate reading processes that are more or less important to comprehension. They do, however, indicate characteristics that are associated with particular texts and items and, as such, provide insight into factors that influence comprehension in particular contexts.

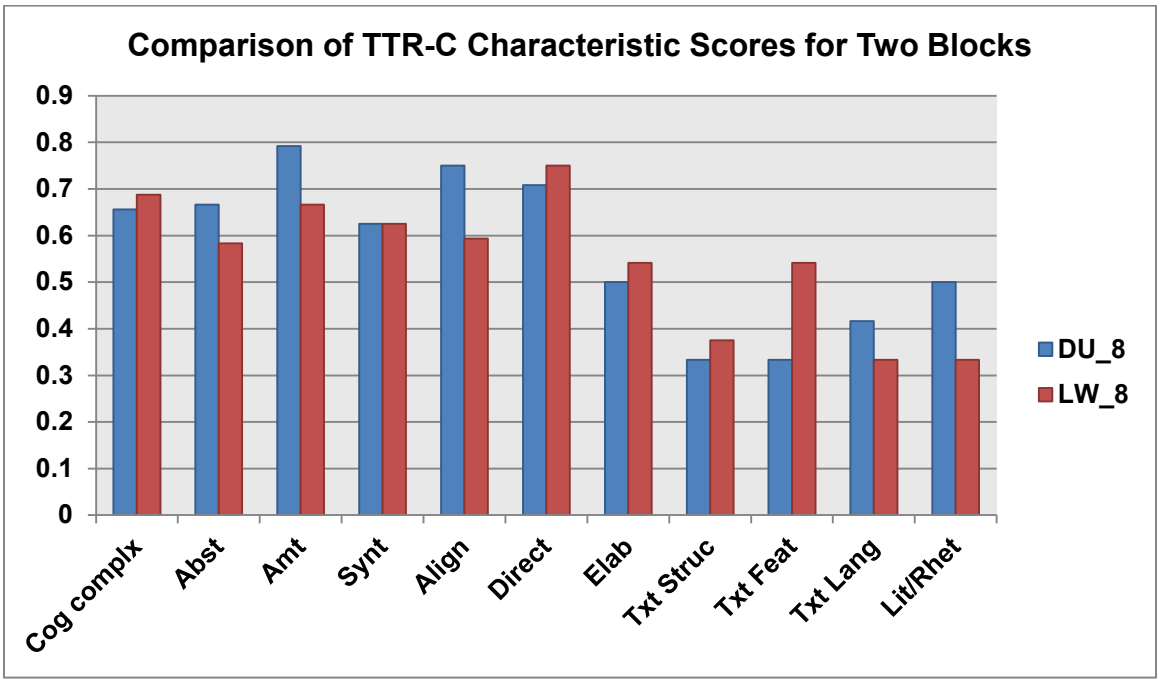
Figure 4. Correlations Between TTR-C Characteristics and Item Difficulty[6]



Similarly, visual inspection of characteristic scores by item within grade-level blocks also revealed a good deal of variability across grades and blocks. Using the grade-level block as the unit of analysis, Figure 5 displays an example of the relative contribution of each characteristic to the overall block difficulty for two of the eighth-grade blocks. Each of the characteristics was standardized to represent scores ranging from .25 to 1.00 to allow comparisons across characteristics. The first block, DU_8, consists of a sophisticated literary passage and eight questions; the second block, LW_8, is a nonfiction piece written in the format of an article, also with eight questions. The average IRT difficulty parameter for each block is identical (midrange difficulty for the grade), yet the pattern of TTR-C characteristics for each block is different and distinctive. Although scores for several characteristics were similar (e.g., Cognitive Complexity, Synthesis, Text Structure), DU_8 was scored higher on the rubric (i.e., more challenging) than LW_8 in the areas of Abstractness, Amount of Text, Alignment, Text Language, and Literary/Rhetorical Style—characteristics consistent with the literary type of text and deep comprehension expected of students in eighth grade. In contrast, characteristics of Stem Directness, Elaboration, and Text Features are rated more challenging on the informational block, LW_8, picking up on the specific nature of this informational text and specific items in the block. These patterns point to the utility of a finer-grained analysis of TTR-C characteristic considerations for understanding comprehension as well as for assessment development and interpretation.

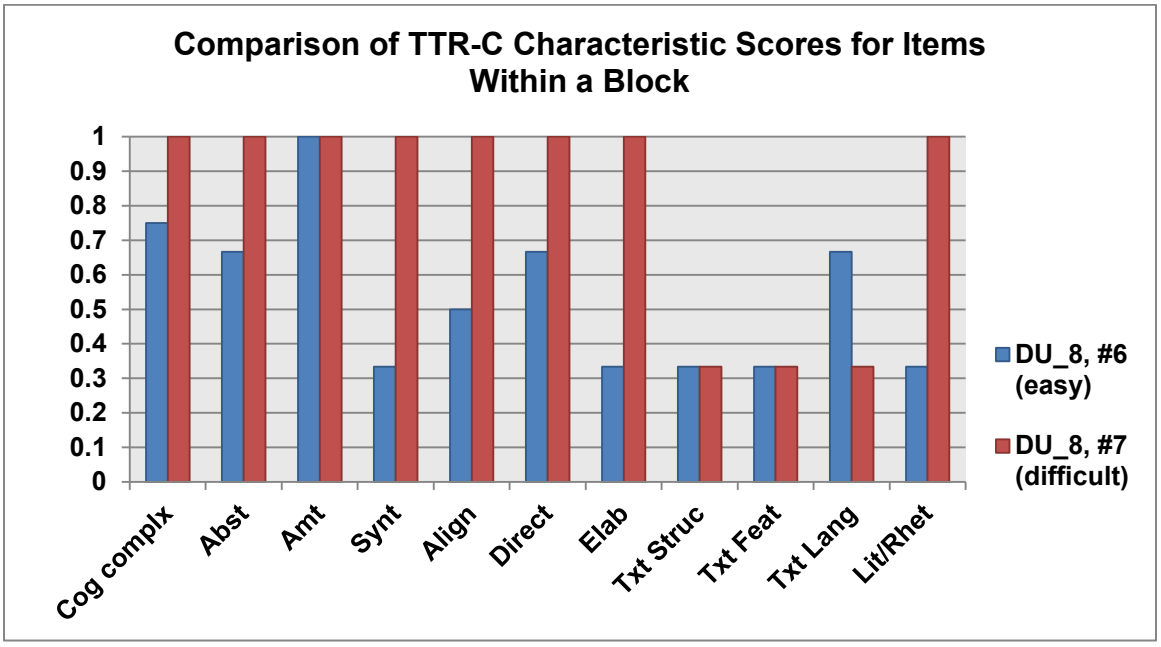
[6] Difficulty calculated as percent correct across blocks within grade level.

Figure 5. TTR-C Characteristic Scores for Eighth-Grade Blocks of Equivalent Difficulty



Analyzing at the next grain size, the item, we examined how TTR-C characteristics might inform specific text and item development and interpretation. Figure 6 demonstrates this approach. Both items are drawn from block DU_8, the eighth-grade literary block displayed in Figure 5. The average difficulty parameter placed this block at midlevel difficulty, and Lexile readability (1040) placed it in grade band 6–8. Both item 6 and item 7 are constructed-response items, yet item 6 was somewhat easy (b=-.6) while item 7 was somewhat difficult (b=.84). The bar graphs provide insight into the characteristics that likely contribute to the difficulty and ease of individual items. It is interesting to note, however, that Text Language is rated more difficult for the easier item than for the more difficult one. Again, this profile suggests that multiple characteristics are at play simultaneously in determining comprehension difficulty.

Figure 6. TTR-C Characteristic Scores for Items Within One Block



Following up on the significant differences in student performance between multiple-choice and constructed-response items (see Item Type in Table 8), we examined the relative challenge of TTR-C characteristics associated with each item type at each grade level. Again, because the blocks (i.e., specific texts and associated items) vary across grade levels and scores of TTR-C characteristics vary across blocks, grade-level comparisons are not useful. However, within grade level, these differences between the characteristics of multiple-choice and constructed-response items provide information about the aspects of comprehension that are being tested with the two item formats. Figures 7–9 provide insight into what may be contributing to the significant differences of student performance at each grade level (all at $p < .01$), beyond item format. For example, at fourth grade (Figure 7), constructed-response items are significantly more challenging in terms of Cognitive Complexity, Abstractness, and Response Elaboration than multiple-choice items; differences in terms of Alignment and Text Features also are worth noting. Although some of these distinctions may seem "logical" given the two item formats, they are not "required" of the formats. Again, the TTR-C characteristics may provide additional considerations for systematic item development within and across item formats.

Figure 7. TTR-C Characteristic Scores for Multiple-Choice and Constructed-Response Items for Grade 4

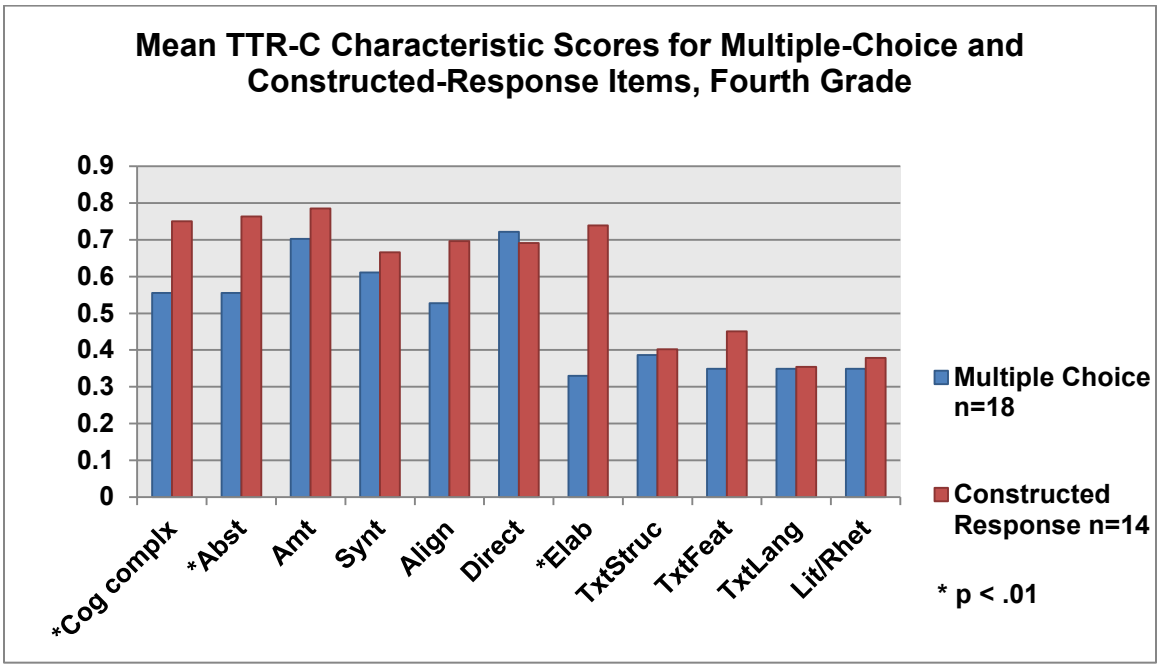


Figure 8. TTR-C Characteristic Scores for Multiple-Choice and Constructed-Response Items for Grade 8

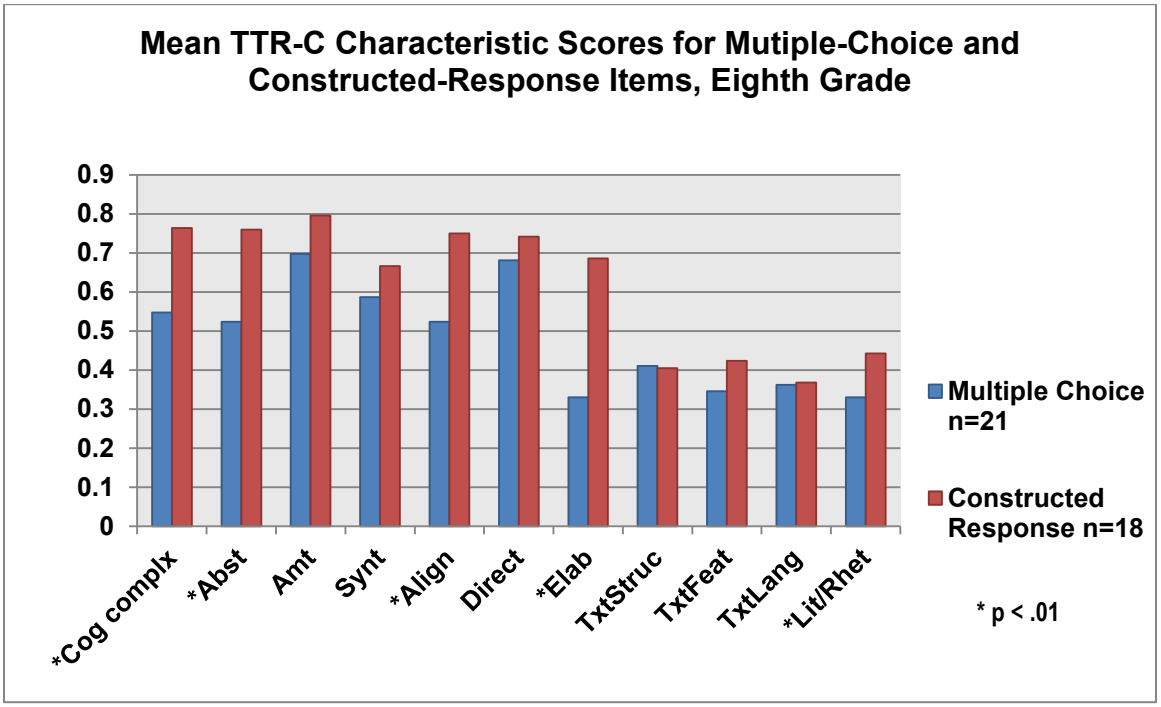
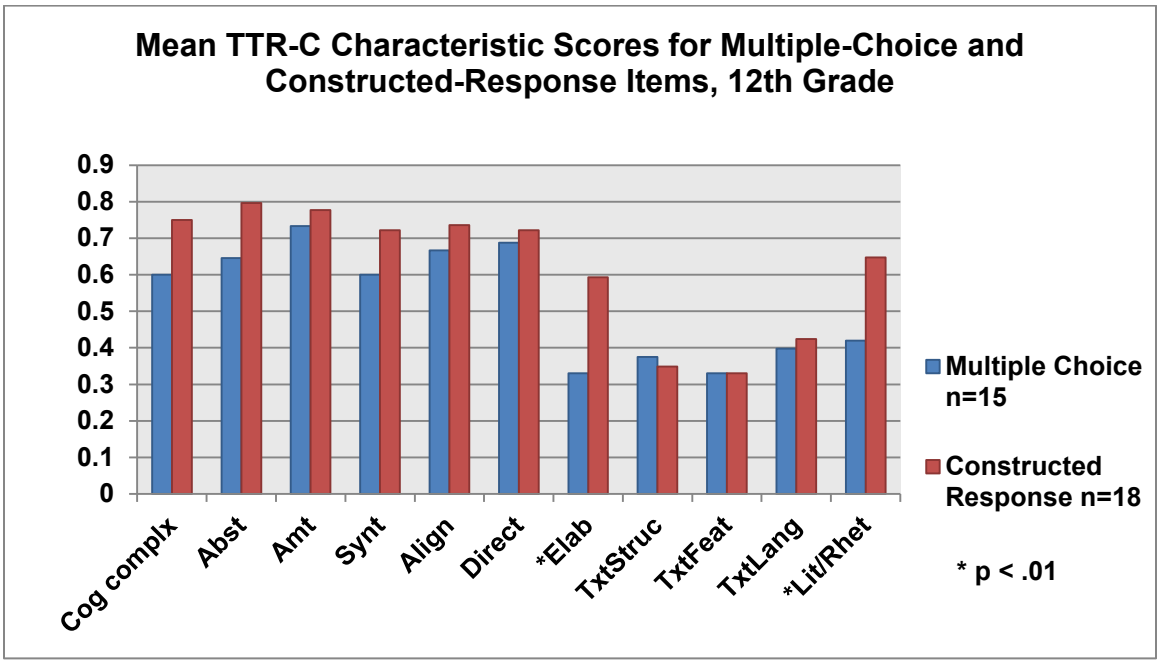


Figure 9. TTR-C Characteristic Scores for Multiple-Choice and Constructed-Response Items for Grade 12



Text-Task-Reader Distractor (TTR-D) Characteristics

The overall finding of significant differences in percent correct for item type (multiple-choice versus constructed-response), as well as differences in correlations and scores of TTR-C characteristics for each item type, led us to additional qualitative examination of the TTR-D characteristics.

Table 9 presents a summary of the TTR-D characteristics for "strong" distractors (>14% response rate) by grade. Each distractor) was coded with as many characteristics (see definitions in Table 5) as applicable, and percentages were calculated to facilitate comparisons within each grade. For example, although there were only 25 "strong distractors" identified at Grade 4, multiple coding with TTR-D characteristics resulted in 67 codes; percentages were calculated using the total number of codes. As with the analysis of TTR-C characteristics, coding of distractors was influenced by text-task-reader interactions as they pertained to specific multiple-choice distractors for specific questions related to specific reading selections.

The sample of blocks and student responses are different at each grade and TTR-D characteristics are not equally distributed across blocks—they are dependent on the nature of the text-task-reader-distractor interaction within each block. We cannot, therefore, compare distractor characteristics across grade level (e.g., there may be fewer distractors representing linguistic/literary features [TTR-D characteristic "c"] at lower grade levels due to the inclusion of fewer complex literary selections and/or items at that grade level). However, overall, the chart suggests that fourth-grade items resulted in more "strong" distractors (i.e., selected by >14% respondents; *N*=25) than did items at eighth (*N*=14) and 12th grades (*N*=9), even though the

number of multiple-choice items at each grade was relatively comparable (fourth=18, eighth=20, 12th=15).

Table 9. Tally of Characteristics of "Strong" Distractors at Grades 4, 8, and 12

Characteristic Code+	a	b	c	d	e	f	g	h	i	j	Total
Grade 4*											
Total	3	9	3	7	10	3	10	3	6	13	67
%	4.5	13.4	4.5	10.5	14.9	4.5	14.9	4.5	9.0	19.4	100
Grade 8^											
Total	4	5	4	7	7	3	9	2	1	9	51
%	7.8	9.8	7.8	13.7	13.7	5.9	17.7	3.92	2.0	17.7	100
Grade 12◆											
Total	1	3	8	3	5	2	5	0	1	5	33
%	3.0	9.1	24.2	9.1	15.2	6.1	15.2	0.0	3.0	15.2	100

+ See Table 5 for characteristic code descriptions.
* *n*=25 strong distractors
^ *n*=14 strong distractors
◆ *n*=9 strong distractors

Examining the patterns within grade level, the analysis illuminates characteristics of distractors in multiple-choice items that make comprehension challenging for students and the ways in which they may "misread" text. The percentage of strong distractors at all three grade levels suggest that students have the most difficulty in three main areas: (1) distinguishing item-relevant events/ideas from others included in the text ("e"), (2) distinguishing between partially correct and fully correct answers ("g"), and (3) distinguishing between what appears as an "obvious" choice and a correct response that requires an inference ("j").

The ability of this scheme to capture TTR-D characteristics may provide test developers with a way to systematically assess and interpret aspects of comprehension that have not been addressed in large-scale assessment. It may be that the difficulty of multiple-choice items is substantially related to the specific aspects of comprehension captured in this TTR-D scheme, and that difficulty can be more predictably manipulated. We explore this further in the following section.

Developmental Trends

We conducted several exploratory analyses using the blocks that were administered at two grade levels (three blocks at Grades 4/8 and two blocks at Grades 8/12) to examine characteristics that might be related to developmental differences across grades. These analyses make it possible to generate hypotheses about developmental differences in comprehension performance.

Item types. Given the overall significant difference between item types, we began by simply comparing percent correct for multiple-choice and constructed- response items in blocks administered to Grades 4/8 and 8/12. Based on the limited number of items in this subset, *t*-tests revealed that constructed-response items were significantly more difficult than multiple-choice items for students at both Grades 4 and 8. Furthermore, although both types of items were significantly more difficult for fourth-grade than eighth-grade students, the differences across item types were similar for both grades; that is, constructed-response items were not disproportionately more difficult at fourth grade than at eighth grade. There were no significant differences between difficulty of item types at either grade level for the blocks administered at Grade 8 or 12 and no differences across Grades 8 and 12.

Distractor analyses. Another source of data for developmental trends comes from the TTR-D characteristic analysis of blocks administered at both fourth and eighth grades. Our analysis indicated that fourth-grade students relied more than eighth-grade students on simple logic or general recall of the text rather than specific text-based information in selecting a multiple-choice answer. An example of this comes from a text about a rescued baby shark, which states explicitly that aquariums have not been able to keep white sharks alive because they will not eat in captivity. In a Locate and Recall item that asks why it has been difficult to keep white sharks in captivity, 36% of the fourth-grade students selected the distractor "they grow too quickly" and only 44% selected the explicitly stated, text-based information, "they do not eat." Information about the size of the baby shark and the fact that it grew quickly is mentioned multiple times in the text in comparison to a single mention of white sharks not eating in captivity, suggesting that fourth-grade students were relying on their general recall of the information in the text rather than searching for the precise answer to this question. By comparison, 72% of the eighth-grade students selected the correct response and 20% selected the incorrect distractor about the shark growing quickly. Although the distractor was still attractive to eighth-grade students, the data suggest that a much larger proportion of the students in Grade 8 were attending to the text-based information.

A similar pattern occurred with a literary text administered at fourth and eighth grades where younger students were more attracted to distractors that fit their general "reading" of the text rather than explicit information, or were attracted to concrete answers rather than generalizations. These analyses may suggest that, when questions are aimed at information that is not central to a coherent understanding of the text—an integrated understanding of the major concepts—eighth-grade students may recognize the need to return to the text to extract that information while younger, more naïve readers may simply rely on their overall impressions. Furthermore, these overall impressions may actually be consistent with important information and goals of the text and/or the reader.

Comparisons of distractor choices for blocks administered at both Grades 8 and 12 revealed that eighth-grade students were drawn to specific distractors three times as often as 12th graders for the same questions, resulting in 12 strong distractors at Grade 8 versus three at Grade 12. There was a tendency of eighth graders to choose distractors that included partially correct information and appeared more "obvious"

than the desired inference when responding to questions scored high (more difficult) in cognitive complexity and processing of a large amount of text.

Block profiles. A final source of data on developmental trends comes from qualitative "profiles" of fourth- and eighth-grade students' performance on identical blocks—one informational and one literary. By examining both the TTR-C and TTR-D characteristics for each item, and the percentage of students at each grade level who selected each distractor for multiple-choice items and earned full credit for constructed-response items, developmental trends begin to emerge. In addition, the profiles provide a level of specificity that might be helpful from a test construction as well an instructional perspective. We provide two examples of "profile" building.

The informational block (WH) was rated at 1000L level of difficulty according to Lexile readability (Grades 6–8). The text was notable for its use of flashback in telling about the life of an inventor, and several of the questions required students to make generalizations and inferences across large portions of this text that were not presented chronologically. Overall, fourth-grade students performed less well than eighth-grade students on multiple-choice items that required more complex cognitive processing or processing large amounts of text (often the entire passage) and that provided little or no specificity/cueing to the location or identification of the correct answer in the text. Fourth-grade students were more likely to be drawn to distractors that included common words between the text and distractor, that were based on logical but inaccurate inferences from the text, and that required an inference that might be missed or was not as obvious as other choices. Eighth-grade students did not have difficulty with these aspects of comprehension except in one item that required integrating information from noncontiguous parts of the text and had distractors that included logical (but inaccurate) inferences and/or common language between the distractor and text.

Constructed-response items were decidedly more difficult for fourth-grade students than eighth-grade students, although they also were difficult at Grade 8, with fewer than 50% of eighth graders earning full credit for any of these items. All of the constructed-response items in this block required students to generalize or make inferences on major ideas across the entire text. They were rated more difficult on more TTR-C characteristics than any of the multiple-choice items, suggesting multiple and more complex challenges. In addition to the TTR-C characteristics of Cognitive Complexity and Amount of Text, which also were present in multiple-choice items, constructed-response items were consistently rated more difficult in Abstractness and Synthesis, and frequently required more elaborated responses for students to earn full credit.

The second profile is based on analyses of a literary block (LN) rated at 800L (Grades 4–5), according to Lexile readability, and administered to both fourth- and eighth-grade students. The story involves the experiences of children from an immigrant family as they visit the home country of their parents for the first time. Abstract themes of culture and identity are notable in this story and are the focus of half of the comprehension items.

Among the blocks we analyzed, this LN block was one of the most difficult for fourth graders and one of the easiest for eighth graders (average 48% correct versus 73% correct, respectively). All five of the multiple-choice questions were rated as difficult on Cognitive Complexity and Abstractness, and all five had correct responses that required either a generalization about, or an understanding of the concepts of identity and culture—concepts that were implied but not defined in the text. In one of the five multiple-choice questions, all three of the distractors attracted 14% or more of the responses at Grade 4, while none of them did so at Grade 8. This question asked the students why a particular paragraph was important to the story, and the correct response used the term "identity," although the text does not include this word.

Overall, fourth-grade students reading this block were about twice as likely as eighth-grade students to choose multiple-choice distractors that included specific language from the text, an actual idea from the text, or partially correct information. Eighth-grade students were less likely to be drawn to this surface-level similarity between question and text and gravitate toward more inferential responses. For example, the correct answer to one question was a generalization embedded among a number of concrete details within the text. At Grade 4, more students selected the distractor about a concrete event from the text than the correct response, even though it was a close paraphrase of the text. The distractor based on a concrete event was still attractive to students at Grade 8, but approximately two and a half times more students at Grade 8 selected the correct answer. In general, Grade 4 students appear to be more likely to select answers that are concrete, especially in response to items that are abstract or cognitively complex and call for generalizations or understanding of abstract concepts.

Responses to constructed-response questions for this block showed differences between the grades that were similar to the differences seen in the informational block. Overall, eighth-grade students earned twice as many total points as fourth-grade students for these multiple-point items—items that were rated difficult in the areas of Cognitive Complexity and Amount of Text. These items also were rated difficult on characteristics of Alignment, Directness, and Elaboration, indicating that they did not use language directly from the text or cue students to particular parts of the text, but they often required students to support their answers with information from the text.

These exemplar profiles and analyses, drawn from identical passages and items administered at two grade levels, provide a window into empirically derived descriptors of comprehension development. By replicating this approach with other cross-grade blocks, developmental trends and potential achievement-level descriptors may emerge that could inform assessment development and interpretation.

Summary and Potential Implications

Summary

We began this study in an effort to identify characteristics reflecting text-task-reader interactions that are related to comprehension difficulty. We also were interested in how the identified characteristics were related to a variety of other block- and item-level factors in predicting item difficulty. At the block level, we were interested in the

relationships between the TTR-C characteristics and quantitative measures of text difficulty (Lexile and TextEvaluator), text genre (Literary and Informational), and grade level (4, 8, and 12). At the item level, we were interested in relationships between TTR-C characteristics and item type (multiple-choice and constructed-response) and cognitive targets (Locate/Recall, Integrate/Interpret, and Critique/Evaluate).

The results indicated that the 11 TTR-C characteristics we defined in this study were related to comprehension (item) difficulty—the rubric used to score the characteristics provided a reliable metric associated with item difficulty. Although seven of the TTR-C characteristics were correlated enough to comprise a single factor when analyzed across all grade levels, further analyses also revealed a great deal of variability across these characteristics for individual items within a block at a particular grade level, as well as across blocks and grade levels. TTR-C characteristics appear to be most related to difficulty based on the specific text-task-reader interactions required by the text and items within each block. In contrast to the results for the TTR-C characteristics, the quantitative measures of text difficulty (i.e., Lexile and TextEvaluator) were not related to comprehension difficulty, suggesting that it is unlikely that comprehension difficulty can be predicted using traditional readability measures without taking into consideration the demands of the reading tasks associated with any given reading event—be it an assessment or an instructional activity.

The results for the other two block-level variables indicated that grade level was significantly related to item difficulty, but text genre was not. Grade 4 blocks were relatively more difficult for their audience than blocks at either Grade 8 or Grade 12, a finding that is consistent with a long-held observation about NAEP. The lack of difference between item difficulties based on literary versus informational texts may be due to a combination of factors, including the restricted sample used for this study, and indicates the challenges that NAEP faces in making full use of purpose or genre qualities in text selection and/or item development (i.e., characteristics associated with text format, text structure, literary/rhetorical style) (see Wixson, Valencia, Murphy, & Phillips, 2013).

Item-level factors in this study included item type and cognitive target in addition to the TTR-C characteristics. Results of the HLM analysis indicated that constructed-response questions were more difficult than multiple-choice questions overall. There were significant differences in scores for many (six of 11) of the TTR-C characteristics by item type, with constructed-response items scoring higher (more challenging). This suggests that characteristics other than item format may be influencing performance—that the items may be tapping different aspects of comprehension. TTR-C characteristics may help unpack these contributions to constructed-response difficulty (and multiple-choice ease) across grades. In addition to these general trends, qualitative profiles conducted at each grade level show evidence of variability in TTR-C characteristics both within and across item types. These TTR-C characteristics are not unique to particular item types and may provide additional considerations for systematic item development to more fully explore facets of reading comprehension tested by different item formats. Added to these TTR-C characteristics, examination of distractor characteristics (TTR-D) also may help to explain differences across item formats and allow developers to construct

item distractors more systematically. Three categories of distractors seem to be particularly rich starting points: (1) distinguishing item-relevant events/ideas from others included in the text, (2) distinguishing partially correct from fully correct answers, and (3) distinguishing answers requiring inferences implied in the text from answers requiring text-based literal information.

With regard to cognitive target, the results suggest that Locate/Recall and Integrate/Interpret items are significantly different than Critique/Evaluate items, but trend toward not being significantly different from each other. Although this last finding may be somewhat unexpected, the qualitative analyses of both TTR-C and TTR-D characteristics illuminate differences that may not be picked up by these cognitive target labels. More specifically, seven of the 11 TTR-C characteristic scores were significantly different across all three cognitive targets, while others showed less consistent patterns. These may provide additional specificity for cognitive target definitions.

Developmental trends were explored on a limited subset of blocks administered across two grade levels. Overall, data from Grade 8/12 blocks were similar across grades while data from Grade 4/8 blocks highlighted developmental differences worthy of further investigation. Based on qualitative block profiles, fourth graders demonstrated more difficulty than eighth graders with questions that were characterized by higher levels of cognitive complexity, use of larger amounts of text, abstractness, and little cueing in the question stem. Fourth graders were drawn to distractors that included logical, rather than text-based information; words from the text; and concrete information rather than inferences or generalizations. These types of grade-level differences were not apparent in the sample of Grade 8/12 blocks analyzed in this study.

Overall, when block- and item-level factors were analyzed together, they accounted for close to 60% of the variance in item difficulty. The results of the full model indicated clearly that both sets of factors were significantly related to item difficulty, with the item-level factors, most notably TTR-C characteristics, accounting for a substantial amount of variance beyond the block-level factors. The additional qualitative analyses highlighted the variable and interactive nature of TTR-C and TTR-D characteristics themselves at the grade, block, and item levels. When considered simultaneously, these characteristics provide rich descriptions of aspects of reading comprehension at work in particular text-task-reader contexts. It is apparent that high levels of challenge in some characteristics may be compensated by relative ease of others and that the aggregate of several challenging characteristics, either TTR-C or TTR-D, contribute to greater difficulty. In sum, these characteristics may have the potential for informing the interpretation of performance on reading assessments and for test development—selecting reading passages as well as constructing items and predicting their difficulty. Below we explore possibilities in these two areas.

Potential Implications

Interpretation of Readers' Performance on Reading Assessments

Describing the text-task-reader interactions characterizing a particular assessment "event" aids in the interpretation of readers' performance. For NAEP, this potentially adds richer description to the item maps and achievement-level descriptors that are used to aid interpretation.

Item maps. NAEP currently provides item maps for the reading assessment that describe the items that "anchor" the achievement levels. The descriptions of these items found on the item maps are typically very general and rarely make reference to the text. As a result, it is often difficult to understand why different items with similar descriptions appear at multiple achievement levels. Examples of some of the item descriptions for the 2015 NAEP reading assessment at each achievement level for Grade 4 are as follows:

> DESCRIPTIONS OF ITEMS ANCHORING AT **BASIC** LEVEL FOR GRADE 4, INCLUDING ITEM TYPE AND COGNITIVE TARGET
>
> Informational Items
> -Recognize main purpose of an informational text (I/I, MC)
>
> Literary Items
> -Infer and recognize main problem faced by story character (I/I, MC)
>
> DESCRIPTIONS OF ITEMS ANCHORING AT **PROFICIENT** LEVEL FOR GRADE 4, INCLUDING ITEM TYPE AND COGNITIVE TARGET
>
> Informational Items
> -Interpret information to describe steps in a process (I/I, CR)
>
> Literary Items
> -Recognize main way author presents information about a biographical character (C/E, MC)
>
> DESCRIPTIONS OF ITEMS ANCHORING AT **ADVANCED** LEVEL FOR GRADE 4, INCLUDING ITEM TYPE AND COGNITIVE TARGET
>
> Informational Items
> -Make a text-based inference to recognize reason for action (I/I, MC)
>
> Literary Items
> -Infer character trait from story details to provide description (I/I, CR)

Using the descriptions of the TTR-C and TTR-D characteristics to augment the current item descriptions could provide a more nuanced description of what students are able to do at different achievement levels. Hypothetically, the fourth-grade Basic, Informational, I/I, MC item described as "Recognize main purpose of an informational text" could read "Recognize main purpose of an informational text

when it is explicitly stated in the text and the key is unambiguously the best choice." Similarly, the fourth-grade Proficient, Literary, C/E, MC that reads "Recognize main way author presents information about a biographical character" might read "Recognize main way author presents informational about a biographical character in a passage with clear text features that signal its organization." Finally, the fourth-grade Advanced, Informational, I/I, MC item that is described as "Make a text-based inference to recognize reason for action" might say "Make a text-based inference to recognize reason for action in the form of a generalization from reading an entire passage, distinguishing it from one or more distractors that include explicit information from the text."

Achievement levels. The NAEP achievement-level descriptors for *Basic, Proficient,* and *Advanced* levels within each grade level also could benefit from the inclusion of information from the TTR-C and TTR-D characteristics. These descriptions are derived from the items anchoring at each level within a grade and, as seen from the descriptions in the item maps, provide minimal attention to the interactions between the items and the texts. For example, the current descriptor for the Basic level at Grade 4 reads:

> Fourth-grade students performing at the *Basic* level should be able to locate relevant information, make simple inferences, and use their understanding of the text to identify details that support a given interpretation or conclusion. Students should be able to interpret the meaning of a word as it is used in the text.
>
> When reading **literary** texts such as fiction, poetry, and literary nonfiction, fourth-grade students performing at the *Basic* level should be able to make simple inferences about characters, events, plot, and setting. They should be able to identify a problem in a story and relevant information that supports an interpretation of a text.
>
> When reading **informational** texts, such as articles and excerpts from books, fourth-grade students performing at the *Basic* level should be able to identify the main purpose and an explicitly stated main idea, as well gather information from various parts of a text to provide supporting information.

Hypothetically, this description would include more information about both the nature of the literary and informational texts in which students at this level were able to perform these tasks, and the characteristics of the items on which they performed well with regard to the text-task-reader interactions. It is well understood that not all texts of a particular genre are comparable in their complexity, so descriptions of texts require more nuanced information. Based on this study, it is most likely that fourth-grade students at the *Basic* level were able to demonstrate the abilities listed in the achievement-level description when the text needed to answer a question correctly was fairly straightforward and concrete and there were no TTR-C text characteristics that made the item challenging (i.e., structure, language, features, rhetorical devices). It also is likely that items provided good cueing as to the location and specificity of the response and did not rely on synthesizing large amounts of text. Furthermore, it is likely that the text/item set did not emphasize abstract themes or generalizations.

Although an appropriate level of specificity would need to be determined for item maps and achievement levels, our point here is that both TTR-C and TTR-D characteristics

could help communicate more clearly the conditions under which students at different levels are performing, and the variables and reading processes that impact student understanding. As a result, they might provide additional information that could inform curriculum and instruction and serve as a check on the foci of assessment.

Block Development and "Managing" Difficulty

A notable finding of this study was that different TTR-C characteristics played significant roles in the difficulty of blocks and items within blocks. Furthermore, these characteristics interacted with each other, and with TTR-D characteristics, to produce more or less challenging items. For example, one of the literary blocks administered at Grades 8 and 12 contained a fictional selection written in a sophisticated literary style, using mature vocabulary and figurative language. Two of the items in this block required an understanding of this style of writing and were rated as difficult on the Text Language characteristic, in contrast to the items in most of the other blocks. With the same scores on Text Language, however, these two items had different scores on the other characteristics, resulting in one item being easy and the other one average for eighth grade. Another question in this block asked about the author's style, producing "high" scores on the Literary/Rhetorical devices and Amount of Text characteristics, again in contrast to the items in many of the other blocks we analyzed, and resulting in a very difficult item. Similar examples exist for informational blocks, especially related to text features such as charts or diagrams included in the item set (or not) that are central to understanding.

Overall, both the TTR-C and TTR-D characteristics and associated rubrics may prove useful in block and item development for NAEP by providing a finer-grained framework for understanding and assessing comprehension. Based on our findings, we suggest the following areas for further exploration:

- Use the TTR-C characteristics as a "blueprint" for item development or a "check" on the complexity of current items. The characteristics may be useful in manipulating difficulty and examining the complexity of the reading process tapped by a set of questions associated with a specific passage. These characteristics also may be a useful addition to passage selection criteria (e.g., more/fewer texts with text features, sophisticated text structure, literary/rhetorical devices). Comparisons of these blueprints and criteria within and across grades would help to document, more specifically, what is being assessed.
- Use the TTR-D characteristics to assist in distractor development. Explore if particular distractors are more or less likely to appeal to readers at various achievement levels. Explore if distractors coded with multiple characteristics are systematically more difficult than others that include fewer characteristics.
- Use the characteristics to understand the general finding that constructed-response items are more difficult than multiple-choice items. Examine the extent to which the item types are assessing similar or different TTR-C characteristics and the role that TTR-D distractor characteristics play. This might be particularly informative in understanding the role of new technology and scenario-based items.
- Use the TTR-C and TTR-D characteristics to analyze cognitive target classification. These may help identify why particular Locate/Recall items are just

as difficult as Integrate/Interpret items (or vice versa) and may help refine cognitive target definitions.

- Conduct further studies to determine how TTR-C and TTR-D might inform developmental trends and processes. Explore how particular characteristics or a combination of characteristics may be more or less influential on difficulty across grades.

The qualitative analyses of TTR-C and TTR-D characteristics also caused us to question the practice of using cross-grade blocks—especially at Grades 4 and 8. The issues with this practice were most evident in blocks containing texts with abstract themes or big ideas and/or using sophisticated language or text structures, which were particularly difficult at Grade 4. The items in these blocks often require students to understand the more complex features of these texts. Perhaps the difficulty of these blocks could be managed by only asking questions about the more "literal" ideas or information in these texts, but this would likely result in a set of items that miss the essence of a particular text. As it stands, however, it appears that cross-grade blocks are too challenging and possibly developmentally inappropriate at the lower of the two grade levels at which they are administered, and insufficiently challenging at the higher of the two grade levels.

Limitations

There are a number of limitations of this exploratory study. The first is the small sample size—four to five blocks per grade level and approximately eight items per block. This sample size certainly limits generalizability. Perhaps more limiting is that we had several cross-grade blocks that may have skewed the findings. Our choice of cross-grade blocks was originally intended to allow for analysis of developmental differences, holding constant the text and items across grades. Although the cross-grade blocks provided useful data to address these issues qualitatively, they likely confounded the quantitative analyses, especially the HLM findings. We suggest that the study be replicated with additional blocks at each grade level and without cross-grade blocks in the quantitative analyses.

A second limitation concerns replicability, both in terms of the representativeness of the "domains" of texts and items tested by NAEP (as noted above) and the in-depth scoring process required to analyze the text-item interactions using the rubric developed for this study. Moving forward, we suggest that the rubric be fine-tuned for use with a larger group of scorers and, as we have done in this study, reliability indices be examined. Along these same lines, we suggest that an abbreviated rubric might be used, eliminating factors that have limited relation to item difficulty.

A third limitation is the way in which text-task-reader characteristics for multiple-choice distractors were handled in this analysis. Although there has been a good deal of research on generic issues regarding distractor development and functioning, our efforts were designed to go beyond those by specifying the text-task-reader interactions that were likely to make some distractors appealing for naïve readers. Although we were able to identify characteristics of distractors that were selected by a comparatively large percentage (>14%) of students at each grade level, there were

too few of these distractors to include in the statistical analyses. Therefore, these factors were only considered in the qualitative analyses.

Finally, beyond the scope of this study, we plan to continue to work on ways in which insights about text-task-reader interactions might be used to more clearly and explicitly describe readers' performance. We imagine there may be ways in which NAEP item maps or achievement-level descriptors might be expanded; there also may be a place for this type of rich description in documentation accompanying released items and blocks as a way to build professional development for teachers.

In conclusion, this study, similar to the work of Kirsch (2001), offers a framework that can be used both for developing the tasks to measure reading as well as for understanding the meaning of what is being reported with respect to the comparative reading proficiencies of students at different grade levels. The framework identifies a set of characteristics that have been shown to predict performance on a variety of reading tasks. Collectively, the characteristics provide a means for moving away from interpreting assessment results in terms of discrete items or a single number, and toward identifying performance in rich detail. Equally important, this highly situated, multidimensional, interactive view of comprehension has a potential role in shaping our thinking about how comprehension of complex text is taught, developed, and assessed.

References

Alexander, P. (2007). Bridging cognition and socioculturalism within conceptual change research: Unnecessary foray or unachievable feat? *Educational Psychologist, 42*, 1.

Alexander, P., & Jetton, T. (2000). Learning from text: A multidimensional and developmental perspective. In M. Kamil, P. Mosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 285–310). Mahwah, NJ: Lawrence Erlbaum.

Armbruster, B. B., & Anderson, T. H. (1985). Frames: Structures for informative text. In D. H. Jonassen (Ed.), *The technology of text* (Vol. 2, pp. 90–104). Englewood Cliffs, NJ: Educational Technology Publications.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Bruce, B. C., Osborn, J., & Commeyras, M. (1994). The content and curricular validity of the 1992 NAEP reading framework. In R. Glaser & R. L. Linn (Eds.), *The trial state assessment: Prospects and realities: Background studies* (pp. 187–216). Stanford, CA: National Academy of Education.

Campbell, J. R., & Donahue, P. L. (1997). *Students selecting stories: The effects of choice in reading assessment.* Washington, DC: National Center for Education Statistics. Available from https://nces.ed.gov/nationsreportcard/pdf/main1994/97491.pdf

Campbell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2001). *Framework and specifications for PIRLS Assessment 2001*. Chestnut Hill, MA: Boston College, Lynch School of Education, PIRLS International Study Center.

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika*, *9*, 185–97.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, *3*, 499–545.

Goldman, S. R., & Rakestraw, J. A. (2000) Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III). Mahwah, NJ: Erlbaum.

Graesser, A. C., McNamara, D. S., & Kulokowich, J. M. (2011). Coh-metrix: Providing multilevel analysis of text characteristics. *Educational Researcher, 40*(5), 223–234.

Guthrie, J. T., Wigfield, A., & VonSecker, C. (2000). Effects of integrated instruction on motivation and strategy use in reading. *Journal of Educational Psychology, 92*(2), 331–341.

Johnston, P. (1984, April). *Assessment in reading: A Vygotskian perspective*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kirsch, I. (2001, December). *The International Adult Literacy Survey (IALS): Understanding what was measured*. Research Report RR-01-25. Princeton, NJ: Educational Testing Service.

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.). *Handbook of reading research* (pp. 681–744). New York, NY: Longman.

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

National Assessment Governing Board. (2015). *Reading Framework for the 2015 National Assessment of Educational Progress*. Washington, DC: Author.

National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010). *Common Core State Standards for English Language Arts*. Washington DC: Authors.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Retrieved from http://www.ccsso.org/documents/2012/measures%20oftext%20difficulty_final.2012.pdf

Ogut, B., Dogan, E., Ndege, M., & Hummel, S. (2010, September). *Predicting difficulty of NAEP reading items with Coh-Metrix factors and readability Indices* (Study Report). Washington, DC: NAEP Education Statistics Services Institute.

Organisation for Economic Co-operation and Development. (2000). *Measuring student knowledge and skill: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy*. Paris, France: Author

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods*, *40*, 1001–1015.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Spear-Swerling, L. (2004). A road map for understanding reading disability and other reading problems: Origins, prevention, and intervention. In R. B. Ruddell & N. Unrau, J. (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 517–573). Newark: DE: International Reading Association.

Stenner, A. J., & Burdick, D. S. (1997). *The objective measurement of reading comprehension*. Durham, NC: MetaMetrics, Inc.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences, 4th Edition* (p. 394). Mahwah, NJ: Lawrence Erlbaum Associates.

Valencia, S. W., & Pearson, P. D. (1986). *Reading assessment initiatives in the state of Illinois, 1985–1986*. Springfield: IL: State Board of Education.

Valencia, S. W., Pearson, P. D., & Wixson, K. K. (2011, February). *Assessing and tracking progress in reading comprehension: The search for keystone elements in college and career readiness*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal, 115(*2), 270–289.

White, S. (2011). *Understanding adult functional literacy*. New York, NY: Routledge.

Wixson, K. K., Valencia, S. W., Murphy, S., & Phillips, G. W. (2013). Study of NAEP reading and writing frameworks and assessments in relation to the Common Core State Standards in English Language Arts. In F. Stancavage, G. W. Bohrnstedt, S. Rosenberg, & T. Zhang (Eds.), *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress*. A publication of the NAEP Validity Studies Panel. San Mateo, CA: American Institutes for Research.

Appendix A

Content Expert Consultants

Karen Wixson (University of North Carolina at Greensboro)

Carol Lee (Northwestern University)

P. David Pearson (University of California, Berkeley),

Charles Peters (University of Michigan, Ann Arbor)

Cynthia Shanahan (University of Illinois at Chicago)

Sheila Valencia (University of Washington-Seattle)

Appendix B

Rubric for Text-Task-Reader Comprehension (TTR-C) Characteristics

FACTOR	1	2	3	4	5
COGNITIVE PROCESSES involved in arriving at correct answer* Degree of cognitive complexity to go from stem to text-based inference(s) back to the correct answer Note: This includes reasoning needed to eliminate distractors.	Predominately Locate/Recall or simple inference within a paragraph; no/little inferential reasoning required	One or two inferences that require more processing than Level 1, but are still fairly obvious or "accessible" to get from stem to text evidence, and back to key/rubric; may include generalizations	Involves drawing conclusions, generalizations, synthesis, and analysis without evaluation	Involves evaluation that may include analysis, application, and critiquing; judgment	Inferences requiring consideration of information from multiple aspects of EACH OF TWO texts
ABSTRACTNESS of text content in relation to key/rubric or process of abstract reasoning to get to key	Concrete content in the stem, text, and key/rubric; little or no abstract reasoning needed to move among stem, text, and key/rubric	A mixture of concrete and abstract content or reasoning to move among stem, text, and key/rubric	Predominantly abstract content or reasoning to move among stem, text, and key/rubric		
(Minimum) AMOUNT of text/source of inference needed for correct answer (Note: Some CR correct responses require more text than others.) Assuming NO prereading, how much text would students need to read to answer correctly? Not including the elimination of distractors; questions with best or most require reading more than one part.	A single sentence or between contiguous sentences within one paragraph	Two to three contiguous paragraphs	More than three contiguous paragraphs; or general understanding of entire selection; two or more noncontiguous sentences or paragraphs (separated by more than one paragraph)	Across two texts	

FACTOR	1	2	3	4	5
SYNTHESIS Process of pulling together information or understanding in the text or to go from text to correct answer	Minimal synthesis Information relevant to answering question correctly is explicit or implicit in a single paragraph or short amount of continuous text (e.g., dialogue).	Information relevant to answering question correctly requires pulling together information/ understanding from more than one paragraph; Could be multiple Locate and Recall (not in same paragraph).	The "whole" of what needs to be understood is more than the sum of the parts; requires inferences at multiple places in the text and then bringing together to get to correct answer.	Same as 2 or 3 but from two different texts	
LANGUAGE ALIGNMENT between text and stem and key(in MC) or rubric (in CR) that leads to correct answer (Factor does NOT include distractors; only correct or full-credit response.)	Identical language or everyday synonyms among stem, key, and text that are helpful to getting correct answer	Some common or paraphrased language among stem, key, and text that is helpful to getting correct answer	Some common or paraphrased language among stem, key, and text, but is not very helpful to getting correct answer	No common or paraphrased language among stem, key, and text	
STEM DIRECTNESS How much specificity or cueing the item stem provides to the location or identification of the correct answer – Cueing location (e.g., at the beginning of the story, at the fair, in the section on whales) –Specificity versus generalizations called for in the stem (e.g., How did the character respond? What happened when? versus the best, most important, mainly)	Clear, concrete cueing or specificity leading to location or identification of correct answer in text	Some cueing or specificity leading to location or identification of correct answer in text	Very little cueing leading to location or specificity of correct answer in text		

FACTOR	1	2	3	4	5
ELABORATION OF RESPONSE (multiple-choice key full-credit constructed response) May be present in both multiple choice and constructed response (e.g., He was persistent because he knew he had a good idea; She was brave—two examples in the story are…)	One idea	Two ideas	More than two ideas required		
TEXT STRUCTURE Genre/Organizational Pattern (e.g., clear narrative structure, flashbacks, point/counterpoint) Exposition—evidence/claims, cause/effect, general listing of information, cohesion, coherence, scientific discourse	Text needed for correct answer doesn't rely on understanding any aspect of text structure	Text needed for correct answer relies on some understanding of text structure	Text needed for correct answer is dependent on an understanding of text structure		
TEXT FEATURES (e.g., headers, illustrations, graphics, maps, sidebars, quotations)	Text needed for correct answer doesn't rely on understanding any aspect of text features	Text needed for correct answer relies on some understanding of text features	Text needed for correct answer is dependent on an understanding of text features		
TEXT LANGUAGE needed for correct answer Includes vocabulary, concrete/abstract concepts, syntax, colloquial language, figures of speech, dialect, and so on	Text needed for correct answer characterized by concrete language, familiar vocabulary, and syntax; Text language NOT LIKELY a factor in getting the correct answer.	Text needed for correct answer characterized by fairly concrete language, although includes some unusual syntax/dialect and sophisticated vocabulary; Text language IS SOMEWHAT a factor in getting the correct answer.	Text needed for correct answer characterized by some vague language, unusual syntax/dialect/ colloquialism or unfamiliar vocabulary; Text language IS a factor in getting the correct answer.		

FACTOR	1	2	3	4	5
LITERARY & RHETORICAL STYLE (e.g., metaphor, irony, symbolism, humor, suspense, characterization, knowledge of literary terms, theme-embedded questions, persuasive language)	Text needed for correct answer does not require understanding of any literary or rhetorical features.	Text needed for correct answer requires some understanding of literary or rhetorical features.	Text needed for correct answer is dependent on a deep understanding of literary or rhetorical features.		

*The term “correct answer” throughout this rubric refers to the correct choice in multiple-choice items or the full-credit response in the rubric for extended-response items.