Invitational Research Symposium on
Through-Course
Summative Assessments

ASSESSING AND TRACKING PROGRESS IN READING COMPREHENSION: THE SEARCH FOR KEYSTONE ELEMENTS IN COLLEGE AND CAREER READINESS

Sheila W. Valencia
University of Washington, Seattle

P. David Pearson
University of California, Berkeley

Karen K. Wixson
University of North Carolina, Greensboro

April 2011

Center for K – 12 Assessment
& Performance Management at ETS

Assessing and Tracking Progress in Reading Comprehension: The Search for Keystone Elements in College and Career Readiness

Sheila W. Valencia
University of Washington, Seattle

P. David Pearson
University of California, Berkeley

Karen K. Wixson
University of North Carolina, Greensboro

The purpose of this paper is to propose a process, a model, and a research agenda for creating both learning progressions and related measures of reading comprehension that align with the English Language Arts Common Core State Standards (ELA-CCSS; National Governors Association Center for Best Practices & Council of Chief State School Officers [NGA Center & CCSSO], 2010). The hope, indeed the expectation, is that recommendations we make in this paper will assist the two state consortia—the Smarter Balanced Assessment Consortium (SBAC, 2010) and the Partnership Assessment for College and Career Readiness (PARCC, 2010)—that have been charged with the responsibility of developing assessments that will measure the capacity of students, teachers, and schools to achieve the CCSS. The assessments under development by PARCC and SBAC are intended to provide useful information to American teachers, schools, and policy makers who aspire to ensure high degrees of college and career readiness among high school graduates.

To set the stage for our recommendations, we begin with a brief review of the proposals for through-course/interim assessments put forward by the two consortia—PARCC (2010) and SBAC (2010). We also take a brief look at the ELA-CCSS (NGA Center & CCSSO, 2010). We then describe and adapt an assessment development process used by Kirsch (2001, 2003) in developing reading assessments for the National Assessment of Adult Literacy (NAAL) and the International Assessment of Adult Literacy (IALS) to guide our efforts to define, organize, and operationalize the construct of reading comprehension for assessment purposes. Finally, we make recommendations about a research agenda, both near and long term, that we might undertake as an assessment community to validate constructs, formats, items, scales, and targets.

Our experiences in reading assessment development over a quarter century, refined by a review of recent research and policy initiatives, lead us to emphasize the significance of taking on this research effort for reading comprehension assessment. Because of the key role that reading comprehension plays in the acquisition of disciplinary knowledge, getting reading comprehension assessment (and instruction) right may be our most important milestone on the pathway to preparing students for success in higher education and the workplace.

PARCC, SBAC, and ELA-CCSS: The Policy Context for Assessment Development

First we considered what the PARCC (2010) and SBAC (2010) have specifically proposed with regard to through-course/interim assessments and learning progressions. We also examined the content of the ELA-CCSS (NGA Center & CCSSO, 2010) and various background documents underlying them.

PARCC

The PARCC consortium (PARCC, 2010) proposed, for each grade level tested, an assessment system that includes three *mini-summative* through-course assessments to be given at approximately equal intervals throughout the year and a comprehensive end-of-year

summative assessment. The results of the through-course and end-of-year assessments would be combined to calculate annual scores in each subject for each student. The intent is to distribute through-course assessments throughout the school year, so that assessment of learning can take place closer in time to when key skills and concepts are taught, making it possible for states to provide teachers with actionable information (information that might shape curricula and instruction) more frequently. The design is such that, together, the components address the full range of the Common Core State Standards (CCSS).

Observing that current interim or benchmark assessments often focus too much on low-level content, the PARCC (2010) design promises to correct this situation by administering high-quality through-course assessments that reflect the best kind of classroom instruction and student work and, consequently, can contribute to decisions about student, educator, school and state performance against the CCSS. The PARCC design signals what good instruction should look like by providing rich and rigorous performance tasks that model the kinds of activities and assignments that teachers should incorporate into their classrooms throughout the year.

In both English language arts (ELA)/literacy and mathematics in grades 3–12, students will take focused through-course assessments after roughly 25% and 50% of instructional time; at the 75% time point, they will participate in an extended and engaging performance-based task. The first two through-course components are designed to measure the most fundamental capacity essential to achieving college and career readiness according to the CCSS: the ability to read increasingly complex texts, draw evidence from them, draw logical conclusions, and present their analyses in writing. These focused assessments offer opportunities early in the school year to signal whether students are on track to readiness. For the third through-course task in ELA/literacy, students will be given extended time to identify or read relevant research materials and compose written essays based on them. Afterwards, students will publicly present the results of that research and writing to their classmates, answering questions or

engaging in debate, so that teachers can assess students' speaking and listening skills using a common rubric. The end-of-year component will build on high-quality, authentic texts at the appropriate level of complexity and will sample a range of cognitive demands, with a bias toward tapping deeper into student depth of knowledge.

SBAC

The SBAC (2010) consortium proposed a comprehensively designed assessment system that will incorporate a required statewide summative assessment, along with two types of optional assessments and tools designed to inform instruction and help students understand where they are in their learning as the year progresses: (a) interim assessments used to track students' learning progress at key points during the year, and (b) a variety of formative tools, processes, and practices for teachers to use to understand what students are and are not learning, so they can adjust instruction accordingly.

The SBAC believes that summative assessments by themselves are insufficient to drive positive change in teaching and learning. Thus, the SBAC (2010) argued that interim and formative assessments are required to promote the teaching and learning required by the CCSS. To that end, interim and formative assessments will be developed and implemented directly under the purview of the consortium—not simply adopted from external sources. Grounded in cognitive development theory about how learning progresses across grades and how competence develops over time, the assessments will (a) work in concert with the summative assessment, (b) allow for more innovative and fine-grained measurement of student progress toward the CCSS, and (c) provide diagnostic information that can help tailor instruction and guide students in their own learning efforts.

The results of the optional interim assessments will not be used for accountability purposes, but will contribute to the summative information. Toward this end, the items on the interim assessments will mirror those on the end-of-year, comprehensive summative assessment, and the interim measures will be reported on the same scale as the summative

assessment. In so doing, the interim assessments have the capacity to become a kind of early warning system for districts, teachers, and students about the progress of individuals and groups toward the summative goals. The SBAC also expects to develop interpretative guides, using the publicly released interim assessment items and performance events to illustrate how the SBAC assessments are manifestations of the CCSS.

Because interim assessments are meant to drive instructional interventions more directly than their summative counterparts, items will be built directly around the concept of learning progressions and the interim assessments will focus on fewer concepts, each probed deeply. The plan calls for reports that can provide more direct student-level information at a finer grain size. Learning progressions are key to the SBAC model; they are to be empirically validated descriptions of how learning typically unfolds within a curricular domain or area of knowledge and skill (Darling-Hammond & Pecheone, 2010). The interim assessments will require the specification of the learning progressions in English language arts and mathematics as a first step in the development process. Once identified, these learning progressions will be mapped to the CCSS, and the evidence-based model again will be applied to determine the knowledge and/or skills a student must demonstrate to show mastery of the steps in the learning progression. By design, a set of items developed to measure learning progressions for the purposes of the interim assessments will be deep in terms of content coverage for each content cluster in the CCSS—a larger number of items will be used to measure small, incremental differences in what students know and can do. Item clusters will be developed that can hone in on students' precise level of understanding of those linked pieces of knowledge and/or demonstration of skills that constitute a progression. The entire array of items and clusters will be organized and driven by a computer adaptive assessment system that is designed to maximize efficiency, precision, and validity. It is efficient because it gets to each student's level of competence on the underlying skill progression as fast as possible. It is precise because it offers the largest sample of items in and around that level of competence for each

student. And because it zooms in on each student's *zone of competence*, it provides deeper and more valid information about each student's status on the key skill progressions that define the domain being assessed. That's the logic of the design, at least.

ELA-CCSS

The full title of the ELA-CCSS is *Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects* (NGA Center & CCSSO, 2010). The specification of subject areas in the title of the ELA-CCSS is the first indication that these standards will be different than most state standards in the areas of ELA. The ELA-CCSS provides an integrated, or at least a highly articulated, view of the areas within the English language arts—reading, writing, speaking/listening, and language. This integrated view is applied to two domains—literature and informational text—for reading and writing at K–5. The standards for grades 6–12 are first organized by ELA and then subject matter to distinguish which standards are the responsibility of the ELA teacher and which ought to be addressed by subject area teachers. Within ELA, the organization is similar to that of the K–5 standards—that is, all four areas of the language arts with reading and writing broken down by literature and informational. In contrast, the subject area sections address only reading and writing, and these areas are broken down according to history/social studies and science/technical subjects.

The integrated view of ELA presented by the CCSS contrasts sharply with the heavy emphasis that has been placed on *reading* as an independent entity in recent years, almost to the exclusion of other areas of the language arts and other subject areas in the school curriculum. When reading is part of an integrated model, the emphasis changes dramatically from the "big 5" of National Reading Panel fame (National Institute of Child Health and Human Development [NICHD], 2000), which have dominated reading for the last decade or more—phonemic awareness, phonics, fluency, vocabulary, and comprehension. Within the ELA-CCSS, phonemic awareness, phonics, and fluency are addressed primarily in the foundational skills addendum to the K–5 standards. Vocabulary is highlighted in the language strand, and

comprehension, alongside composition, is emphasized throughout. Add to this the emphasis on reading and writing in the disciplines of history and science at 6–12, and the ELA-CCSS represents a major shift from a dominant emphasis on decoding to a consistent emphasis on comprehension of and learning with text.

Taking Stock

Whether we are talking about through-course summative assessments or optional interim assessments, there is a need to identify learning progressions. As noted in a recent report (Doorey, 2011), the determination of the subset of skills and concepts to be emphasized in these assessments requires that we identify within each content area and grade level the keystone topics or cognitive targets for which deep mastery is necessary—and highly predictive of—readiness for college and a career (or, for earlier grades, predictive of later performance on key cognitive targets). What are those skills and concepts, what level of mastery do they require, and in what sequence (if one such sequence exists)? Studies will need to be carried out to gain deeper understanding than we currently have to support these decisions (Doorey, 2011). This is essentially what this paper articulates in the area of reading comprehension.

Our Guiding Model of Reading Comprehension Assessment Development

The work of Kirsch (2001, 2003) has not only shaped our thinking about the entire test development process, but it has dramatically impacted our thoughts about how to address the key first step in the process—developing a model of reading assessment, which, among other things, would allow us to determine the factors that ought to be used to shape text selection and item development. And it is our strong recommendation that test developers use a process similar to the one used by Kirsch (2001, 2003) and his colleagues with the IALS and NAAL as they develop reading assessments for the new era of Common Core State Standards. Then and only then can we be assured that we have an assessment that is as well grounded in a sound

conceptual model of reading comprehension as it is in a sound psychometric model. The steps in the model used by Kirsch for both the NAAL and the IALS include the steps depicted in Figure 1.

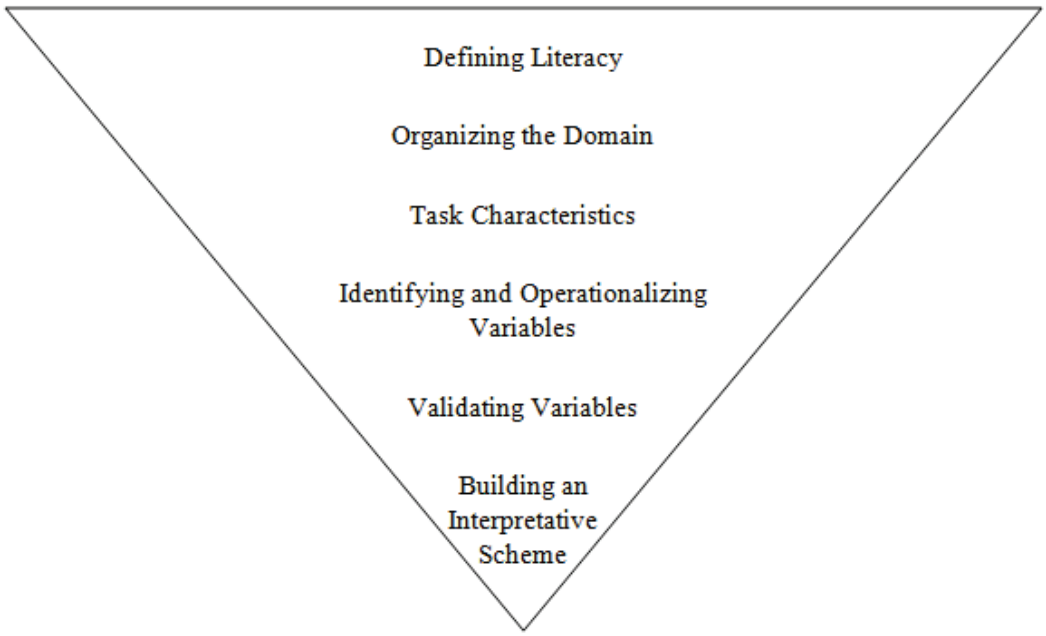

Figure 1. Steps in developing a literacy test. From The International Adult Literacy Survey (IALS): Understanding What Was Measured (ETS Research Rep. No. RR-01-25) by I. S. Kirsch, 2001, Princeton, NJ: ETS. Copyright 2001 by Educational Testing Service. Reprinted with permission.

Step 1, *defining literacy*, begins the process. In using this model for the IALS, Kirsch completed Step 1 more in the domain of everyday use than the more limited domain of schooling: "Literacy is using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential" (Kirsch, 2001, p. 6). In conceptualizing a reading assessment for the era of the Common Core State Standards, which is the work envisioned by PARCC and SBAC, we begin at the beginning or, if you prefer, at the *end*

of the process, with a clear statement of what we mean by reading comprehension. As will become readily apparent to readers of this essay, this has been our most difficult and complex, as well as our most important, challenge.

Step 2, *organizing the domain*, is both a conceptual and a psychometric concern because the number of categories one selects in creating an infrastructure for a construct like reading comprehension will influence both the number and nature of items designed to measure each key element in the domain. And this step, of course, is guided by the definition, research, and theoretical model selected in Step 1.

Step 3, *specifying task characteristics*, takes the process to the next logical step in specificity. Kirsch, 2001), in discussing this step, noted its importance in the work of Almond and Mislevy (1998), who noted that variables can be used to limit scope, construct tasks, control the distribution of tasks across forms, report performance, or interpret proficiencies. Thus, Step 3 hearkens back to Steps 1 and 2 and anticipates Steps 4–6, even to the point of interpretation.

Step 4, *identifying and operationalizing variables*, is the heart of assessment development, for it is in this step that the tasks are made real. For Kirsch (2001), this turned out to be a rather long and complex matrix of variables falling into three groups: content/context (of adult reading), text/materials (the stuff of reading), and processes/strategies (processes needed to either identify or construct the correct response from the information available).

Step 5, *validating variables*, is a decidedly psychometric process, conducted in the best spirit of research and development, in which there is an iterative process of small scale engineering-like design studies, test pilots, and field trials in order to establish the validity of the constructs that were used in Steps 1 and 2 to set the process in motion.

Step 6, *building an interpretive scheme*, builds on the prior decisions by providing a useful means for exploring the progression of demands across each of the scales and what

scores along a particular scale mean. Thus, it contributes to the construct validity of inferences based on scores from a measure.

In one sense, there is nothing special about the six steps that Kirsch (2001) outlined; they are, in broad strokes, part and parcel of the rigorous processes that have guided test development for decades. But in another sense, the process Kirsch employed is unique in the degree to which it is accountable to the highest of implementation standards at each step along the way. Many efforts skip lightly over certain steps, especially Steps 1–3, and, as a result, they provide us psychometrically sound but conceptually suspect assessment tools. And the evidence of the value of the IALS and NAAL work is in the rich body of validity evidence that they have accumulated in documenting the constructs that underlie both IALS and NAAL.

Applying the Kirsch Development Process to New Comprehension Assessments

By applying the principles of the Kirsch (2001, 2003) model to our task of assisting efforts such as PARCC and SBAC, we hope to contribute to an assessment development effort that can do justice to the possibilities generated by the new ELA-CCSS.

Step 1: Defining Reading Comprehension—Sources

Defining reading comprehension is a major component of our approach to assessment development. As such, we will devote disproportionate attention to Step 1 of Kirsch's framework in developing a model for assessments that can do justice to the ELA-CCSS.

To reach a definition, we consulted the cognitive, curricular, and pedagogical research and policy documents that bear on the assessment of reading comprehension. Specifically, we have focused our attention on the following resources:

- The largely cognitive account of the nature of reading comprehension, along with its instruction and assessment, as detailed in the federally sponsored report of the RAND Commission (RAND Reading Study Group, 2002).

- The cognitive literature that has defined our conceptualization of how people learn (e.g., Bransford, Brown, & Cocking, 1999) and cognitive examinations of how they learn from text, including the important and often overlooked areas of motivation, as reflected in constructs such as interest, incentives, choice, and monitoring one's own learning (e.g., Alexander, 2003; Alexander & Jetton, 2000).
- The instantiation of these key cognitive concepts, bolstered by many from literature and the teaching of language arts, in the 2010 ELA-CCSS developed in a national effort led by the Council of Chief State School Officers and the National Governors Association (NGA Center & CCSSO, 2010).

The perspectives elaborated in each of these three sources were then filtered through the assessment development steps outlined by Kirsch (2001, 2003) to create the current document. These sources are important for our effort because each provides a piece of the puzzle we must solve to define reading comprehension. Only with a credible definition in place will we be able to build a research agenda that has the integrity to stand up to the theoretical, scholarly, and policy forces that will scrutinize any assessments that emerge from this effort. Only by taking all of these views into account can we create an agenda that reflects our best, research based knowledge of learning, reading comprehension, measurement, and policy issues.

A significant milestone: The RAND report. In 1999, the RAND Corporation was charged by the National Academy of Science with the task of defining reading comprehension and setting a research agenda to better understand the nature of reading comprehension and how better to teach and assess it. In the process, the group responsible for the effort provided the literacy field with a consensus view of reading comprehension that reflected the important cognitive and sociocultural work on reading over the previous three decades. The blue ribbon panel that RAND appointed defined reading comprehension as "…the process of simultaneously extracting and constructing meaning through interaction and involvement with written

language. We use the words *extracting* and *constructing* to emphasize both the importance and the insufficiency of the text as a determinant of reading comprehension" (RAND Reading Study Group, 2002, p. 11).

RAND Reading Study Group (2002) went on to suggest that comprehension entails three primary elements:

- The *reader* who is doing the comprehending.
- The *text* that is to be comprehended.
- The *activity* in which comprehension is a part. (p. 11)

Finally, the panel acknowledged that the act of comprehension that entails these three elements always occurs in a sociocultural context "that shapes and is shaped by the reader and that interacts with each of the three elements" (RAND Reading Study Group, 2002, p. 11). The panel depicted this set of relationships what has become a widely used graphic (see Figure 2), which serves as an icon of our collective knowledge about reading comprehension as a conceptual, pedagogical, and psychometric phenomenon. The RAND definition emphasizes the salience of both the text (extracting meaning) and the reader (constructing reading) through interaction (that's the activity) with written language.

Reader. The RAND report (RAND Reading Study Group, 2002) posited a central role for the reader in the comprehension process. Included in reader factors are cognitive capacities (e.g., attention, memory, critical analytic ability, inferencing, visualization ability), motivation (a purpose for reading, an interest in the content being read, self-efficacy as a reader), and various types of knowledge (vocabulary, domain and topic knowledge, linguistic and discourse knowledge, and knowledge of specific comprehension strategies). Although the RAND Reading Study Group (2002) outlined the implications of the definition for content area reading instruction, disciplinary knowledge (e.g., the nature of the work of history, physics, or mathematics) is not as salient as topical knowledge (e.g., photosynthesis, density, the nature of revolutions, or rational numbers).

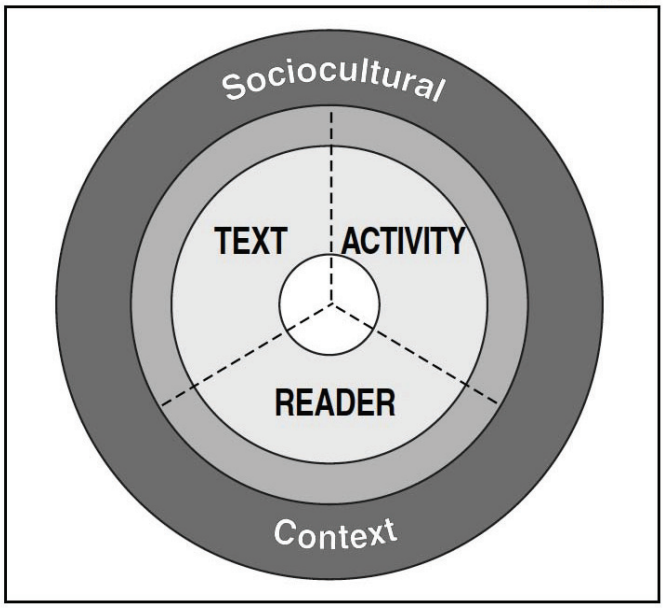

Figure 2. The RAND model of reading comprehension. From Reading for Understanding: Toward an R&D Program in Reading Comprehension, by RAND Reading Study Group (p. 12), 2002, Santa Monica, CA: RAND. Copyright 2002 by RAND. Reprinted with permission.

Text. The RAND report provided an elaborate account of text features that impact comprehension, including:

- Genre (elaborated as narration, description, exposition, and persuasion).
- Structure: the *organization* of the text (how authors organize text within and across paragraphs and short segments: rhetorical structures such as problem-solution, compare-contrast, sequential steps/events) and the *overall coherence* (degree of similarity of ideas from one sentence to the next).
- Media forms (e.g., textbooks, multimedia, advertisements, hypertext, Internet).
- Sentence difficulty, including vocabulary (familiar to rare words or concrete to abstract words) and syntax (simple to complex grammatical structures).

- Content (the transparency/obscurity of the ideas represented, including the cultural assumptions that are made by authors about the knowledge and cultural practices readers bring to the text).
- Engagingness, operationalized as the degree to which the text will appeal to groups of readers who bring particular personal interests with them.

Activity. Reader and text comprise two pillars of the comprehension process, but the comprehension edifice cannot stand without considering the activity pillar, that is, the particular cognitive behaviors we take as evidence of comprehension. We are likely to draw different conclusions about an individual's reading competence based on the activity we examine to determine how or how well she reads. Do we examine decoding, fluency, vocabulary knowledge, comprehension, or critical analysis of the text? Each of these activities reveals a different facet of reading. And any one of them varies according to the purpose for which one is reading; decoding might be more deliberate and more carefully monitored when reading the directions for using ant poison than getting the gist of the morning paper. The RAND report includes three broad categories under activity:

- Purpose/task (e.g. skimming for gist, studying to retain information, or savoring the beauty of language).
- Operations to process the text (e.g., decoding, higher-level linguistic and semantic processing, or self-monitoring for comprehension).
- Outcomes of performing activity (e.g., an increase in knowledge, a solution to real-world problem, or increased engagement with text).

Sociocultural context. And, finally, all of these components—reader, text, and activity—get played out in contexts that both shape and are shaped by the components. Context includes the location and associated expectations found in settings such as classroom-learning environments (including organizational grouping, inclusion of technology, or availability of materials) and various workplace settings (time frame, pressure, colleagues) affect the

development of comprehension abilities. The amount of support or interference in such settings plays a major role in understanding and learning from text.

In addition, context includes the identity, motivation, and background of both the reader and the writer, as well as the purpose for engaging in specific reading-related tasks and the nature of the text itself. It makes a difference, for example, whether one is studying for a test, reading for the gist of the argument, or just reading for pleasure—and it is almost always the context that determines purpose. Readers talk back to contexts and, in turn, shape them. Without reader resistance to norms, we would never, in literary theory, have moved from writerly (where the text dominates) to readerlyy (where the constructive power of readers dominate) views of literary interpretation. And even something as simple as unearthing the author's basic argument depends upon the discipline within which one is reading. While there are structural and logical similarities in the nature of arguments across academic disciplines, what counts as evidence or a valid assumption are different in chemistry, mathematics, history, and literary theory.

The interactions among the components. The notion of interaction is crucial to comprehension, for interaction rejects fixed views of text meaning in favor of views that are constructed in accordance with the particular constellation of these components that hold for any given act of comprehension. For example, a given reader—imagine Henry as an eighth grader who scores at the fourth-grade level on a standardized test—might look like a struggling reader in his history class when reading the state adopted textbook. But give him a book on a topic he knows about and has a passion for, basketball for example, and he'll look like a 10th-grade reader. Or change the purpose and/or context of the reading from studying for a test to working on a community project that might benefit his neighborhood, and his comprehension of relevant books and articles might rise to his grade level. Or change the setting and the accountability from one in which everyone reads it on his or her own and writes an individual summary to a small group competing with other small groups for the best book-jacket snippet

for the novel the class has just read, and his comprehension might look different still. Or frontload the chapter on Egyptian cultural contributions with an engaging movie of the archeological discoveries of ancient Egypt, and his comprehension might improve dramatically. All this is by way of saying that neither reading ability nor reading disability is entirely "beneath the skin and between the ears" (Mehan, 1993, p. 241). Competence varies as a function of all the reader, text, activity, and contextual factors we have discussed thus far. To account for the influence of these variations, we need both pedagogical routines and assessments that acknowledge this inherent interaction. Thus, we need to know not just that Henry is an eighth-grade reader, also the reader (skills, strategies, knowledge, and personal interest), text (genre, complexity, quality of explanation, cohesiveness), activity (studying for a test or reading for gist or reading to gather information for another purpose), and contextual (different disciplines, different settings) conditions that might render Henry a more or less able reader. The implications of this highly situated and interactive view of reading are consequential with respect to the ways in which current assessments are developed. Specifically, they suggest that text readability is a function of more than the factors that are associated with traditional or more recent readability formulas (Coh-Metrix, n.d.; Klare, 1984; Lexile, n.d.). Text readability is surely shaped by these text difficulty indicators, but it is also shaped by the interaction of text factors with reader, activity, and contextual variables.

We can create valid assessments only if our tests reflect a reasonable range of variation on these factors, which, to anticipate where we are headed, is the agenda for the later parts of this paper. Specifically, we will identify what we will define as Text-Task Scenarios that could guide a more complete, comprehensive, and differentiated assessment of reading on the pathway to success in college or career settings after K–12 schooling.

Other cognitive perspectives. The RAND report presents a decidedly cognitive view of reading comprehension, with a nod to the influences of the sociocultural context on how we conceptualize, teach, and assess it. But two additional perspectives add depth and breadth to

the picture provided by RAND—the benchmark National Academy of Science report, *How People Learn* (Bransford et al., 1999), which tackled the more general question of learning in all domains, and Alexander's model of disciplinary learning (Alexander, 2003; Alexander & Jetton, 2000), which addressed issues of reading and learning from text within disciplinary settings. The question we asked when examining both of these reports: how do they deepen the knowledge and enhance the model provided by the RAND effort?

How People Learn. Because *How People Learn* (Bransford et al., 1999) focused on learning in general, it said little about the source of the input of information that is transformed, through learning, into knowledge that can be stored in long-term memory; thus, reading, listening, or experiencing are all equally valid potential sources of information that, through learning, can be transformed into knowledge. But it does assign a central role to the prior knowledge that learners bring to any learning opportunity as the most important factor in determining the depth and breadth of learning that will result. In addition to a deep foundation of factual knowledge, learners must have conceptual frameworks (ways of examining the world) that assist in transforming facts into useable knowledge that can be used in inquiry and problem solving.

The learner is the central focus of *How People Learn* (Bransford et al., 1999), with its triadic emphasis on the impact of topical knowledge on learning, the structure of knowledge, and metacognitive monitoring ability. Since these features all reside within the learner, it follows that reader factors play a central role in learning in general and learning from text in particular. The primary focus is on the learner and the important role of the teacher in both building on students' prior knowledge to bridge to new understandings and confronting misconceptions when prior knowledge is inaccurate and misleading.

Bransford et al. (1999) posited an important role for strategy use. Teachers, must help students learn how to monitor understanding, make note of situations in which additional information is required, and determine the degree of consistency between new information

and what is already known. Teachers have to be mindful of analogies that can be drawn to advance understanding and consider alternatives pathways to achieving desired knowledge goals. There is a clear emphasis on strategic learning to acquire new information and solve problems. Bransford et al. made much of the notion of adaptive expertise (i.e., knowing what you know, what you don't, and how to get the information you need). They had precious little to say about text, which is not surprising, since *How People Learn* explicitly acknowledged a range of input modalities in the learning process.

Bransford et al. also acknowledged the importance of activity in recognizing the role of instructional design in establishing (a) an appropriate level of task difficulty (challenging but within reach); (b) purposes within the social context of learning, including collaborative learning; and (c) relevance—helping students see the usefulness of what they are doing/learning and how their learning can impact themselves and others.

Alexander's model of domain learning (MDL). Alexander's MDL perspective (Alexander, 2003; Alexander & Jetton, 2000) does not explicitly define reading or literacy, but it is, by Alexander's own admission, consistent with if not directly influenced by, the RAND report, or, perhaps more accurately, the body of research that also influenced the RAND report. MDL's unique, and from our perspective most important, attribute is its focus on how motivation and affective factors interact with domain knowledge and cognitive strategies to produce learning from text. Alexander (2003) argued that reading and "its development over the lifespan involve both skillful processing and aesthetic response and much more" (p. 48). She went on to argue that

> …reading is an emotional domain, not a coldly cognitive enterprise. Reader motivation and affect are powerful forces in this journey toward competence. We may be able to move individuals into low levels of competence by attending only to their knowledge or strategic needs, but they will not have the personal interest to continue this journey, especially as the demands of text-based learning become greater and the text more complex, less predictable. Creating a stimulating learning environment, selecting texts

> that are exciting or moving, or constructing tasks that are fun or arousing are certainly part of this motivational package. However, educators must work to plant the seeds of individual interest early in students' domain journey because their "motivation from within" will ultimately sustain them. (Alexander, 2003, p. 51)

Her perspective is important for those of us who believe, as we do, that it is impossible for students to achieve the standards outlined in the CCSS for college and career readiness unless students develop the motivational infrastructure (interest, stamina, and self-efficacy, to name a few elements) needed to support cognitive learning. The failure to account for these dispositional factors may, in fact, help to explain the lack of predictive power of earlier tests—they are not tapping the attributes needed to achieve the CCSS. Whether we can assess, or in some other conceptual or statistical manner account for, these affective dispositions, as important as they are, is but one of the vexing questions that confronts educators and psychometricians.

Alexander's MDL (Alexander, 2003; Alexander & Jetton, 2000)includes all of the reader elements outlined in the RAND model (RAND Ready Study Group, 2002), such as various forms of knowledge, a range of strategies, important cognitive processes, and a range of motivational factors. The special developmental focus of Alexander's MDL, however, suggests that knowledge, motivation, and strategies play different roles depending on where students are in their learning. As noted earlier, Alexander (2003)was much more insistent than our other key sources on putting motivational factors front and center in the learning process.

In the papers (Alexander, 2003; Alexander & Jetton, 2000) in which she laid out her MDL, Alexander said little about the nature of text and its role in the comprehension and learning process. The most plausible explanation is that Alexander chose to focus our attention on the processes and products of learning from text rather than the influence of the text input itself into the key processes. We draw this inference because Alexander had much to say about the nature and role of text in other treatments of her model (Alexander & Jetton, 2000).

Alexander's treatment of text in the 2000 piece is highly elaborated and consistent with the RAND approach.

Activity is a central feature of the implicit architecture of Alexander's model (Alexander, 2003; Alexander & Jetton, 2000), and activity is captured in the processes that readers/learners employ in making sense of the data they encounter through experience, be it through reading texts or the world around them. In Alexander's model, activity, defined as a set of both low- and high-level strategies, changes as readers develop competence, as depicted vividly in Figure°3. As readers move from acclimation, a stage of novice ability to learn, to a stage of competence and then proficiency expertise, some sources of knowledge and motivation and types of processing become more, while others become less, salient in learning from text within disciplines.

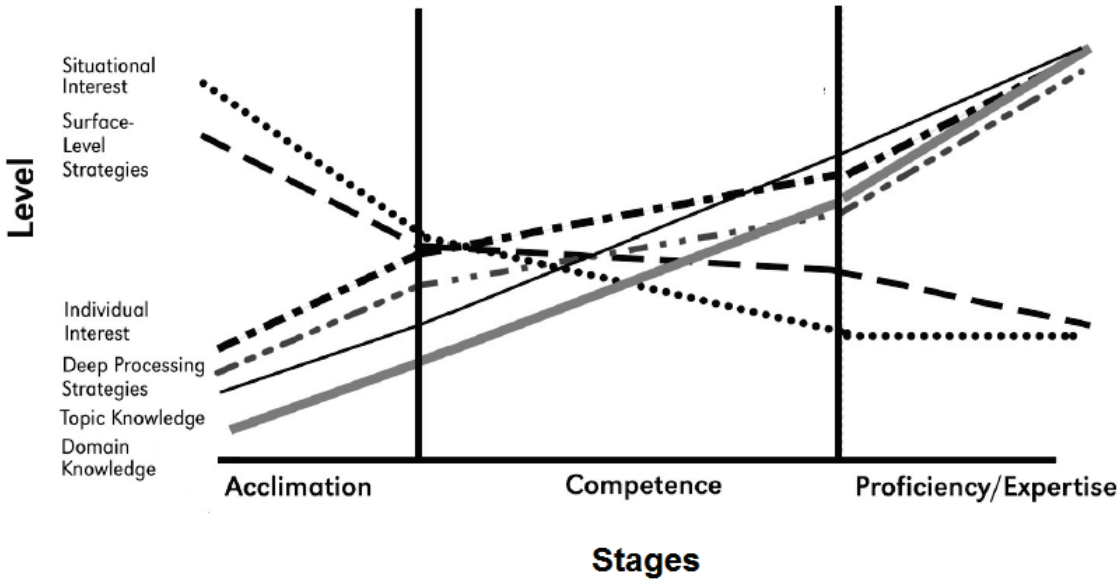

Figure 3. The interplay of knowledge, interest, and strategies across the lifespan. Based on Alexander(2005).

***Value added by cognitive perspectives*.** So, what do we gain from adding the perspectives from *How People Learn* (Bransford et al., 1999) and Alexander's MDL (Alexander, 2003; Alexander & Jetton, 2000? Several things, we think that contribute to a much more elaborated view of the nature and role that cognitive and metacognitive strategies play in comprehension and learning. The RAND report (RAND Reading Study Group, 2002) acknowledged the role of strategies, but the two additional reports add much more detail about what they are and how they impact learning. Thus, they contribute substantially to our conceptualization of reader and activity components of the RAND model. First, from Bransford et al. (1999), we learn that hallmarks of expertise are highly adaptive and flexible strategies and knowledge that is characterized by its utility in solving new problems that require transfer to new situations or, perhaps, even a change in learner purpose. Second, with Alexander's MDL we gain a differential appreciation for the relative salience of knowledge, processes, and motivational factors in learning (as depicted in Figure 3) at different stages of development on the pathway to expertise. Third, from Alexander in particular we gain a richer and more nuanced sense of the role that motivational factors, particularly situational and personal interest, will play in shaping comprehension and learning. Approaches to assessment or instruction that fail to acknowledge motivation will be doomed to inadequate explanations of performance or learning. As we argue in our later section on Text-Task Scenarios, motivation plays a central role in determining the challenge and complexity of texts that students are able to respond to in reading assessments.

The English Language Arts Common Core State Standards. In searching for our definition of reading comprehension, we also used the lens of the ELA-CCSS, again looking for value added to the definition of reading comprehension found in the RAND (2002) report. Our initial finding was that the ELA-CCSS (NGA Center & CCSSO, 2010) did not define literacy, reading, or ELA directly—but it did provide some relevant insights about the fundamental nature of reading and literacy. The closest thing to definitions are statements about the vision

of what it means to be literate in the 21st century (p. 3) and a portrait of what students who are college and career ready in ELA "look like" (p. 7). The vision suggests that to meet the standards, students must

> …readily undertake the close, attentive, reading that is at the heart of understanding and enjoying complex works of literature. They habitually perform the critical reading necessary to pick carefully through the staggering amount of information available today in print and digitally. They actively seek the wide, deep, and thoughtful engagement with high-quality literary and informational texts that builds knowledge, enlarges experience, and broadens world views. They reflexively demonstrate the cogent reasoning and use of evidence that is essential to both private deliberation and responsible citizenship in a democratic republic. (NGA Center & CCSSO, 2010, p. 3)

Designers of the standards (e.g., Coleman, 2010) often refer to this kind of reading as *reading like a detective* (emphasis ours), where the emphasis is on close attention to the textual information in order to piece together a coherent account of what the text means. The portrait of students who meet the standards includes several attributes commonly identified as good reading habits—attributes such as (a) demonstrating independence; (b) building strong content knowledge; (c) responding to the varying demands of audience, task, purpose, and discipline; (d) comprehending as well as critiquing; (e) valuing evidence; (f) using technology and digital media; and (g) understanding other perspectives and cultures.

The Reader in ELA-CCSS. The ELA-CCSS document (NGA Center & CCSSO, 2010) does not directly address most of the reader factors mentioned in the RAND (2002) document. Even so, when it comes to topical knowledge, the document provides a much more explicit account than RAND of the knowledge-comprehension relationship, particularly in the vision statement and portraitof a reader. The ELA-CCSS do attend to vocabulary knowledge and linguistic/discourse knowledge largely through the Language standards, acknowledging a central role for both in both reading and writing performance.

The ELA-CCSS (NGA Center & CCSSO, 2010) do not directly address processes—cognitive abilities, motivation, or experience; in fact, the standards studiously avoid the use of terms such as *strategies* and *processes*. The stance on strategies/processes is that they are instrumental tools, the means by which teachers help students achieve the college and career goals set by the standards. And they offer teachers wide latitude in deciding which to emphasize.

By emphasizing required achievements, the standards (NGA Center & CCSSO, 2010) leave room for teachers, curriculum developers, and states to determine how those goals should be reached and what additional topics should be addressed. Thus, the standards do not mandate such things as a particular writing process or the full range of metacognitive strategies that students may need to monitor and direct their thinking and learning. Teachers are thus free to provide students with whatever tools and knowledge their professional judgment and experience identify as most helpful for meeting the goals set out in the standards.

The text in ELA-CCSS. With one exception, the ELA-CCSS document (NGA Center & CCSSO, 2010) addressed all of the RAND (RAND Reading Study Group, 2002) factors through the College and Career Readiness (CCR) Standard 10—Range of Reading and Level of Text Complexity—as well as the description of text complexity elaborated in Appendix A (NGA Center & CCSSO, 2010). The exception: the ELA-CCSS treated technology as an aid to learning rather than as a distinct type of text requiring different skills to master. The treatment of text in the ELA-CCSS bears strong implications for assessment, especially as test developers decide how to select texts for reading comprehension assessments:

> To build a foundation for college and career readiness, students must read widely and deeply from among a broad range of high-quality, increasingly challenging literary and informational texts. Through extensive reading of stories, dramas, poems, and myths from diverse cultures and different time periods, students gain literary and cultural knowledge as well as familiarity with various text structures and elements. By reading texts in history/social studies, science, and other disciplines, students build a foundation

> of knowledge in these fields that will also give them the background to be better readers in all content areas. Students can only gain this foundation when the curriculum is intentionally and coherently structured to develop rich content knowledge within and across grades. Students also acquire the habits of reading independently and closely, which are essential to their future success ((NGA Center & CCSSO, 2010).

It is worth noting the interplay between text factors and reader factors explicit in this directive as a way of foreshadowing our central theme of interaction among all of the elements in the RAND model (reader, text, activity, and context). Clearly the ELA-CCSS (NGA Center & CCSSO, 2010) expected knowledge to be both the cause (knowledge influences the nature and degree of comprehension) and consequence (students acquire new knowledge through reading comprehension) of text comprehension.

Activity in the ELA-CCSS. Activity is less explicit in the ELA-CCSS (NGA Center & CCSSO, 2010) than in RAND (RAND Reading Study Group, 2002), but it is clearly a part of the fabric of the document. First, it is clear that there are many operations and outcomes in the ELA-CCSS view of reading—everything from close reading to understand the author's meaning to comparing and evaluating different accounts of the same event or theme. Second, in describing the able reader, the ELA-CCSS were quite explicit about the range of operations good readers engage in. They

> …adapt their communication in relation to audience, task, purpose, and discipline. They set and adjust purpose for reading, writing, speaking, listening, and language use as warranted by the task. They appreciate nuances, such as how the composition of an audience should affect tone when speaking and how the connotations of words affect meaning. They also know that different disciplines call for different types of evidence (e.g., documentary evidence in history, experimental evidence in science). (NGA Center & CCSSO, 2010, p. 7)

Context in the ELA-CCSS. The CCSS acknowledges the importance of context largely through the perspective that instruction and, by implication, assessment, must be situated within academic disciplines, even in elementary grades. In our view, this move toward thinking about how each discipline contextualizes the fundamental processes of reading comprehension in its own unique way—and how these differences shape both pedagogy and assessment—is the real genius of the ELA-CCSS (NGA Center & CCSSO, 2010), what sets them apart from previous state and national standards. Even if we eventually learn that the processes used in history are nearly identical to those used in science or literature, having begun the process looking for the unique features that each discipline requires will give us greater confidence in whatever pedagogical and assessment models and tools we develop.

A Definition of Reading Comprehension and Its Assessment as an Interactive Process

Looking across the perspectives that we have examined (Alexander, 2003; Alexander & Jetton, 2000; Bransford et al., 1999; NGA Center & CCSSO, 2010; RAND Reading Group, 2002), we offer the following definition of reading comprehension to frame the next iteration of reading comprehension assessments to accompany the ELA-CCSS:

> Reading comprehension is an interactive and multidimensional process in which students understand, learn from, and use text to accomplish specific purposes in educational, workplace, and everyday settings.

This definition is useful because it emphasizes four important aspects of comprehension as it applies to college and career readiness standards and assessment of reading comprehension more broadly: (a) accuracy (understanding the text); (b) learning (text is one important source of new knowledge in all disciplines and vocations); (c) using information (application is valued in all postsecondary settings, including college, the workplace, and even in the everyday life of being a citizen and consumer); and (d) the complex, multifactor, interactive nature of reading competence. This definition also has strong potential for

generating good 3–12, and especially 9–12, predictors of success in college, careers, and even across the 3–12 continuum (e.g., using grade 6 indicators to predict grade 8 or 9 performance).

Step 2: Organizing the Domain of Reading Comprehension

With the central task of a conceptual definition completed, we turn to the important but problematic work of organizing the domain of reading comprehension; our history of assessment has not served us well in this enterprise. Most often, the quest for an ideal way of organizing the domain of reading comprehension has been to determine the optimal set of subskills in reading. This approach aligns well with commercial reading programs, which universally take this approach to instruction and assessment. Underlying this quest has been the perennial question of the grain size at which we monitor reading progress and offer instruction. For example, we could view comprehension as a unitary construct and assess it with a single score. We could move down a grain size and provide scores for, say, *literal* comprehension of explicitly stated ideas, *inferences* from text to prior knowledge, and *critical* analysis of the ways in which the author portrayed particular ideas in the text. Or we could dig a little deeper and provide estimates of a range of often tested and taught subskills, such as those listed in Table 1, which were used by Frederick Davis in 1930s–1960s to determine the infrastructure of reading comprehension.

The quest for specificity. On the face of it, it would appear that more specificity is always an advantage, especially if one's goal is to provide teachers with highly specific information about how individuals are performing. How else are we to know how to differentiate instruction to bring all students up to the standard we would like all to achieve? However, neither conceptual perspectives nor empirical evidence support a fine grain size approach to the assessment of reading comprehension. First, conceptually, a strong argument can be made that fewer rather than more comprehension components capture the essence of understanding and learning from text—that what we want to promote is the integration of potentially separable, independent skills and processes rather than their separation. The

primary argument for holistic rather than componential conceptualizations of any higher order intellectual performance (text understanding, problem-solving, or applying information gathered from multiple sources to new settings) is that a holistic approach to assessment and pedagogy corresponds much more directly to the context in which we expect students to actually use those higher order skills and processes. In real problem-solving or project-based learning settings, we do not, for example, have a literal comprehension phase, a drawing inferences phase, and a critical reasoning phase. Instead, we ask students to orchestrate a range of these tools in the service of task completion and problem solving. To assess or instruct the alleged components in piecemeal fashion could distract us and our students from the very sort of integrated performances that we desire to promote (see, for example, Returning to Learning, Benjamin et al., 2009). In this more holistic approach, it is still possible to gather diagnostic information that would help instructors alter their teaching techniques to improve performance, but always focused on the component skills and processes as they are enacted in orchestration with other skills and processes in the solution of authentic problems and projects that are central to the course or discipline in which they are embedded.

The evidence. The empirical evidence from a wide range of factor analytic studies of reading comprehension assessment (where the goal is to identify the independent elements of reading comprehension) most often reduces a large set of factors to two or three. For example, Davis (1944; 1968) began with eight factors (see Table 1) but found that word meanings explained the most (32%) unique variance, followed by "drawing inferences from content," with 20% unique variance, which was followed, in order of magnitude, by "structure of the passage," "writerly techniques," and "explicit comprehension." Thus, he concluded that while comprehension was not a unitary factor, nor was it a vast set of componential skills, each acquired independently in some predetermined sequence. In the heyday of factor analytic studies in the 1950s and 1960s, several such endeavors resulted in similar conclusions, namely

that a large number of potentially unique components usually reduced to a much smaller set, often a word factor, a reasoning factor, and an inference factor (see Pearson & Hamm, 2005).

Table 1. Davis's Eight Potential Factors

1. Remembering word meaning	5. Drawing inferences from the content
2. Word meanings in context	6. Recognizing author tone, mood, and purpose
3. Understanding explicitly stated content	7. Recognizing literary techniques
4. Weaving together ideas in the content	8. Following the structure of the content

By the mid 1970s, we witnessed a sharp decline of this rich area of scholarship, with the general view among reading educators being that there were not nearly as many distinct subskills as the available tests and instructional programs of the era would lead one to believe. That consistent and compelling body of evidence would not, however, stop the proliferation of single skill tests, which became even more popular in the 1970s and 1980s as skills management systems, such as Wisconsin Design for Reading Skill Development (e.g., Otto, 1977; Otto & Chester, 1976). The basal readers of the 1970s and 1980s were filled with tests like these for up to 30 different comprehension skills (Johnson & Pearson, 1975). More importantly, they persisted and flourished in the face of many professional critiques of their theoretical and practical efficacy, validity, and utility (Johnston & Pearson, 1975; Valencia & Pearson, 1987).

The professional thirst for more and more subskills abated for about a decade during the mid 1980s through the mid 1990s, when constructivist approaches to teaching and more open approaches to assessment (e.g., performance assessments and portfolios) captured our collective attention. But that era proved to be but a momentary diversion, because it was effectively challenged by the quest to return to the logic of assembling (in instruction) and

assessing (in our officially sanctioned testing programs) separate component skills in the era of No Child Left Behind.

It is worth noting that the empirical evaluations of comprehension assessments in state assessments (see the Illinois work cited in Pearson & Hamm, 2006, for example) and national assessments (e.g., the National Assessment of Educational Progress [NAEP]) have not successfully documented the efficacy of the cognitive infrastructure used to create the assessments. In Illinois, several factor analytic studies revealed a clear trend toward passage topic rather than component processes (such as literal understanding, inferential reasoning, and critical reasoning) as the key factor in explaining how items clustered with one another. In NAEP, it is no accident that performance for the separate stances that guided the development of items from 1992 through 2007 (forming an initial understanding, developing interpretation, reader-text connections, and critical analysis of text) were never used as reporting variables. This is because items on NAEP did not cluster into these categories across passages; that is why NAEP chose to report scales by reading purposes (reading for information, reading to follow directions, reading for literary experience). Not only did the stances fail to emerge from factor analytic studies, they also failed the test of professional judgment. In three separate studies (Bruce, Osborn, & Commeyras, 1994; DeStefano, Pearson, & Afflerbach, 1997; Pearson & DeStefano, 1993), researchers documented the fact that professional reading educators could not reliably classify items according to the stances that the items were ostensibly developed to assess.

Making sense of the evidence. The most compelling interpretation of these data is that, at least for reading comprehension assessment, these cognitive categories implode on one another, that in every inference or evaluation task, there is a bit of literal comprehension, and even literal comprehension can and does involve inference and critique. The other, decidedly complementary, interpretation is that texts (and the knowledge that underlies them) are inherently redundant; thus, there is almost always more than one place to find information

relevant to answering a question, be it literal, inferential, or evaluative in character. And some of those alternative sources invite a more literal approach to getting the right answer, and others, a more inferential or critical path to the same answer. In short, information sources provide great compensatory options and opportunities.

Alternatives to skills-based, cognitive domain organizers. There are alternatives to using questionable subskills or cognitive domains to organize the domain of reading comprehension. We could, for example, organize by context (i.e., in-school academic contexts and non-school work and home contexts) or according to types of materials (literary or informational). Some of these alternative schemes have been used in existing assessments. For example, the NAAL—designed to assess how adults use printed and written information to function adequately at home, in the workplace, and in the community—rejects a definition of literacy as a single ability that one either possesses or lacks. Instead, it defines literacy as task-based, including different types and levels of literacy and corresponding abilities. NAAL measures literacy along three scales that are derived from three types of literacy—prose, document, and quantitative. Each scale comprises the knowledge and skills needed to perform the corresponding assessment tasks. In contrast, other assessments organize the domain around very broad genres or types of texts such as information, literary, and document (e.g., NAEP, many state assessments) or continuous and noncontinuous text (IALS) and report scores according to these scales. Another, yet unexplored, possibility might be to organize the domain around disciplinary fields such as science, social studies, or social sciences to align with what we know about learning from text in the disciplines.

Our examination of the data on failed attempts to organize the domain by skills-based cognitive domains and some alternative frameworks used for test design and reporting leads us to conclude that how the domain is organized is critically important. It shapes the texts that are selected, the tasks and items that accompany each text, the curriculum materials and instructional approaches, and the interpretation of results; directly and indirectly the

organization of the domain helps to define literacy. However, neither data nor theory provides clear direction. We suggest that a preliminary organizational scheme is necessary to create a test blueprint but that scheme, as well as alternatives, must be investigated empirically before test development guidelines and reporting categories are established. Furthermore, we conclude that it is possible, and indeed likely, that both empirical and conceptual decisions about how the domain is organized will change across developmental levels defined by age or grade in school. We deal with this need for empirical studies of these questions in the research agenda that follows.

Summary: Defining the infrastructure of reading comprehension. In conclusion, because reading comprehension is multidimensional, involving many contributing factors to successful understanding, it is clear that more is needed than traditional or even more current readability formulas to determine the difficulty of a text and its associated tasks. In addition, because data do not provide clear evidence of independent comprehension subskills or cognitive categories, organizing the domain to develop assessment tasks is currently simply a matter of convenience and convention. The most important implication of these findings is that we do not possess the research base to establish unassailable a priori learning progressions for reading comprehension; neither cognitive benchmarks nor curricular topics nor text factors can be used alone to determine a sound learning progressions for reading comprehension that will lead to college and career success. Establishing such a progression is part of the research agenda we propose later in this paper—one that finally integrates what we know about all the reader, text, context, and interactive factors that influence how well people comprehend different texts for different purposes.

Steps 3 and 4: Identifying and Operationalizing Variables and Task Characteristics

With the preceding sections on definitions of reading and organization of the domain as background, we propose a model for operationalizing core concepts in reading comprehension assessments. It is our position that reading comprehension assessments must reflect the

complex, dynamic nature of comprehension as enacted by readers while reading texts in real time. Otherwise, assessments cannot serve as models for good instruction or as valid predictors of college and career success.

We propose a unique approach: constructing Text-Task Scenarios as a model for assessment development. More specifically, Text-Task Scenarios assess and require students to engage and work across the many factors that we know influence comprehension as they read a variety of texts to accomplish a range of specific purposes. Although assessments cannot address all the variables that influence comprehension, they can address those we know to be major influences. Therefore, Text-Task Scenarios are designed to tap a variety of purposes for reading, range of texts, and types comprehension tasks, as well as taking into consideration students' interests, prior knowledge, and reading strategies as they read specific texts. In the past, test designers have either fixed text and task variables or they have sampled across them. And they have rarely conditionalized test scores as a function of reader factors, such as interest, knowledge of topic or discipline, or known levels of skill or strategic competence. Text-Task Scenarios do not treat these variables in isolation but as interacting pieces of the comprehension process—and, therefore, of the comprehension assessment process.

In Text-Task Scenarios, reading comprehension performance is defined as readers' ability to engage strategically with a variety of texts and tasks about which their knowledge, expertise, or interest varies substantially. It is this ability to adjust to various demands and one's own skills, dispositions, and knowledge that is the hallmark of competent comprehension. Specifically, text selection includes a range of literary, discipline-specific, out-of-school, and in-school texts. A purpose for reading is set before the student reads each passage, and that purpose is aligned with why people at various developmental levels might authentically read it. Similarly, the comprehension assessment items or performance tasks actually align with the purpose for reading and developmental levels. Indicators of the reader's prior knowledge and interest are taken into account as performance is quantified and qualified across texts and

tasks. For example, a Text-Task Scenario for high school seniors might ask them to read an op-ed article on the pros and cons of judicial appointment versus election to the 9th Circuit Court of Appeals. One follow-up task might ask students to take a position on this debate and use evidence from the article to articulate their position. Performance would be analyzed with consideration of the reader's knowledge and interest level and compared with his or her performance on other Text-Task Scenarios, such as a response to an op-ed piece on teen pregnancy or an informational article on judicial selection. As a result of planned, systematic comparisons such as these, the difficulty of each Text-Task Scenario is empirically established as readers engage in the complex and multifaceted process of reading for understanding. The research agenda outlined below includes representative scenarios of the sort that might help us establish learning progressions within and across grades.

Steps 5 and 6: Validating Variables and Building an Interpretive Framework

This brings us to the heart of the test development process—how we validate the variables we have determined to be valid components of the construct in question, and how we can develop ways of reporting performance on those variables, whether for groups or individuals, that helps us make valid decisions for vital matters such as instructional placement, instructional intervention, or fitness for special programs or privileges. We have concluded that we do not currently have the research understandings required to accomplish Steps 5 and 6 in the Kirsch (2001, 2003) process—that instead what we need are ambitious near- and mid-term research agendas to develop the knowledge and insights required to get it right. Without embarking on this effort, we worry that a decade or two from now, we will be standing at this same juncture, poised to address the same shortcomings and dilemmas that face us today. Just as well we address them now. Thus, we turn to our concluding section—the research agenda we need to accomplish this goal.

A Research Agenda for Validating New Reading Comprehension Assessments and Learning Progressions

The evidence reviewed in this paper leads us to conclude that current reading comprehension assessments do not reflect our best research-based knowledge about reading comprehension, with the result that, as a field, we cannot convincingly identify learning progressions that reflect the complex nature of higher levels of comprehension needed for college and career success. Nevertheless, it is our view that the evidence points the way to improved comprehension assessment through a research agenda that can illuminate learning progressions. We are at a critical choice point in the field. We can either drag on with the same set of constructs and tools that have failed to allow us to define learning progressions for reading in the past, *or* we can admit our shortcomings and work toward a conceptualization of learning progressions that is sufficiently complex to match the nature of comprehension across texts, tasks, dispositions, and domains of knowledge.

We have divided our research agenda into two parts: (a) a short-term agenda that could be undertaken immediately using existing measures and knowledge and (b) a mid-range research agenda of 3–5 years in duration, that would move the field much closer to the type of through-course/interim reading comprehension assessments that more fully reflect the level of reading and thinking required for K–12, college, and career success, as described in the ELA-CCSS.

Short-Term Research Agenda

The short-term research agenda takes a three-pronged approach: an analysis of existing assessments, an evaluation of the predictive validity of existing assessments, and the adaptation of existing assessment materials to a new and more complex conceptualization of comprehension tasks.

Evaluating the match between existing tests and the definition of reading comprehension as an interactive process. The first step is an analysis of existing

comprehension assessments against a definition of reading comprehension as an interactive process that we describe at the beginning of this paper and the Text-Task Scenario model we propose for assessment. The goal of such an analysis would be to identify comprehension assessments that may, with some minor revision, move closer to the desired goal of measuring the complex process of higher level comprehension. At the same time, this analysis could document, more specifically, the definitions and nature of comprehension measured by each assessment, providing useful data to guide schools in test selection and score interpretation.

Predicting future success. The second short-term research activity will determine how well existing grade 3–12 comprehension assessments predict college and career success. Since one of the goals of through-course/interim assessments is to predict future success, this set of studies takes aim at the criterion measures that have traditionally been used as indicators of college and career success. Existing studies have used a range of criterion variables to evaluate the predictive validity of early-warning assessments that could be given in high school. In college settings these include (a) overall grade-point average in college, (b) college retention, (c) grades in particular benchmark courses (e.g., calculus, writing, basic science, and the like); in workplace settings, the list extends to (a) employment earning data, (b) longevity in the workplace, and (c) supervisor satisfaction. All of these criterion variables come with predictable limitations. Rarely have studies used direct measures of the ability to understand, learn from, and use text to accomplish specific purposes in educational, workplace, and everyday settings.

In the last several years, a few measures of comprehension have been developed for adults that come much closer to assessing the definition of complex comprehension required for success (e.g., Collegiate Learning Assessment, NAAL). Using these new assessments as criterion measures would enable us to begin to consider more authentic measures of reading on both sides of the prediction equation—in secondary school and in college and career settings—and such efforts could inform how we should use information from existing measures (e.g., ACT, SAT) that are explicitly designed to predict later success. At the same time as we

explore the best available predictors we can find, we would simultaneously begin the process of developing new models of through-course/interim comprehension assessments more conceptually consistent with our definition of reading comprehension as an interactive process. In essence, this line of research will assure, in the short run, that criterion measures are subjected to as much scrutiny as predictive measures and that they align with expectations outlined in the CCSS for college and career readiness.

Repurposing current assessments. The last item on the short-term research agenda item takes advantage of the pool of well-analyzed, high-quality texts that have already been approved for use on existing comprehension assessments. Because identifying potential reading assessment passages is time consuming and expensive, existing passages may provide an easy starting place for research on new models. This research would involve developing Text-Task Scenarios for existing passages, including attention to systematic variation of texts, tasks, and purposes, as well as investigating existing measures and strategies for assessing reader interest and knowledge. Data from small-scale administration of these prototypes would be used to investigate the interaction of all these variables as they contribute to high levels of comprehension. The results gathered from these rapidly developed adapted assessments would provide key information for conceptually valid developmental trajectories and learning progressions and, ultimately, reporting scales.

Mid-Term Research Agenda

The mid-term research agenda for through-course/interim reading comprehension assessments builds on the short-term agenda by adding a fuller range of texts and tasks, comprehensive pilot testing, and more in-depth psychometric analyses. Specifically, we envision two phases of the research: (a) validating and scaling of Text-Task Scenarios and learning progressions and (b) addressing issues of generalizability head-on.

Systematic development and evaluation of Text-Task Scenarios. First, Text-Task Scenarios must be conceptualized and developed to align with research, theory, and authentic

reading expectations for students in grades 3–12; this is the construct validity question and, as such, is much more than a psychometric issue. Here, we envision a strong role for experts in subject matter (from the various disciplines in which reading is measured), reading theory, and reading curriculum. Visions of what counts as learning with and from text must be based on solid theory, research, and best practice rather than common practice. We envision beginning with three to four touchstone grade levels across the middle and high school years, then working backwards to the elementary levels. It is also essential that a definition and vision of successful college and career reading frame the development of Text-Task Scenarios from the start (close backward mapping is essential here) and that issues related to criterion measures (criterion validity) be considered at the same time. As Shepard (2009) suggested, these new assessments must be piloted in the context of settings in which we know (or at least have good reason to believe) that high quality instruction is being enacted; it is only in such settings that we can conduct the existence proofs required to demonstrate what students can learn and do in the best of circumstances—a critical step in determining learning progressions. Empirical models will also need to be developed to test the contribution and interaction of multiple variables (e.g., text type, reading purpose, comprehension task, disciplinary focus and topic, strategic flexibility, reader interest and knowledge) across developmental levels. Based on these data, learning progressions can be modeled and empirically evaluated.

Consequential validity. In addition to addressing content and criterion validity, we recommend a focus on consequential validity during this phase of research. As outlined in the PARCC proposal, through-course assessments should be designed to signal what good instruction should look like through rich and rigorous performance tasks that model the kinds of activities and assignments that teachers should incorporate into their classrooms throughout the year. Unlike much of the prior research on assessment where consequential validity studies are conducted after an assessment is finalized (and where revision is rendered moot), we

suggest that research on teachers' understanding and use of these new assessments be conducted as part of the development process.

Formative adaptations. In addition to these approaches to consequential validity, research should explore how Text-Task Scenarios can be used by teachers to gather more in-depth diagnostic information to inform instruction; in short, we should evaluate their formative capacity to shape day-to-day, or at least unit-to-unit, adaptations in feedback and scaffolding for individuals and small groups of students (Black & Wiliam, 1998; Shepard, 2000, 2008, 2009). Here, we are interested in exploring a model of assessment that might provide both accountability data (through-course data) as well as instructional data, along the line advocated by Black and Wiliam (1998). As we noted earlier, through-course/interim assessments rarely provide information at a grain size specific enough to inform instruction. However, because Text-Task Scenarios model authentic reading that might happen in a classroom, and because previous research has provided at least partial validity for certain assessment strategies teachers can use to get more in-depth information from these scenarios (e.g., running records, think-aloud protocols), a series of studies should explore this secondary use of through-course/interim assessments to inform instruction in the spirit of formative assessment. Few comprehension assessments have this potential to be used for accountability and, with specific follow-up assessment strategies, to provide more diagnostic information to inform instruction for some students. We can think of no in-principle reason why they cannot or should not be used for both.

***Other psychometric issues*.** As with all assessment development, several types of scaling studies will need to be conducted. Given the interactive model of comprehension that underlies our proposed assessments and the need to empirically establish learning progressions, studies will need to explore multidimensional scaling. In addition, we will need a new generation of factor analyses of a reconceptualized domain of reading comprehension. We need, for example, to assess the stability and co-variation of similarly developed Text-Task Scenarios

across disciplines, as well as across topics within disciplines. Together, these studies could provide a blueprint for comprehension assessment development in the future—how we select texts, tasks, and domain/topical settings and how we develop a model for reporting results that meets both summative and formative needs. Ultimately, both empirical and conceptual data should inform these organizational and operational frames.

Adaptive assessments. We also recommend studies of adaptive and out-of-level testing, with the expectation that these efforts will have strong implications for growth modeling and value-added approaches to program and teacher evaluation. In our view, efforts during the 1980s and 1990s to create rigorous on-grade-level assessments, created a problem for our field. When we adopted the idea that tests designed to measure the mastery of standards at a given grade level (e.g., fourth grade NAEP) should sample from passages appropriate for that grade, we made it difficult to measure performance at the tails of the distribution—either very low or very high performance—validly and reliably; the passages would be incredibly hard for the lowest performers and incredibly easy for the best. This text sampling procedure is to be contrasted with the long-standing practice, among developers of standardized tests, of intentionally sampling passages that represent a somewhat larger range of performance in the population (some that would be easy and some difficult for, say, fourth graders). As a matter of empirical fact, this problem is more pronounced for the lower than the higher tail of the distribution. Pearson, Callahan, and Benson-Griffo (2008), in evaluating the difficulty of fourth grade NAEP passages for a special project on creating more accessible texts and tasks for low-performing fourth graders, found that the NAEP passage selection process results in many more passages with readability levels above (5–8 range) than below (2–3 range) the target range. Further, this trend was especially salient for informational passages. With their well-documented overall poor performance, low-achieving students contribute little information to our estimate of grade level reading achievement in NAEP—and perhaps on other on-grade-level assessments. Translated into practical reporting issues, while we know that 40% of America's

fourth graders read below basic, we don't have enough items in that below basic range to describe what it means to read below basic. And, of course, this problem is particularly acute for students who perennially score low—English language learners, students with disabilities, linguistic and ethnic minorities, and poor students in general. Furthermore, even when low achievers make progress, grade level driven through-course/interim assessments may not be able to detect growth because they don't include passages that students can read. Until research provides models that can be used to document growth of low-achieving students, through-course/interim assessments cannot achieve all of their intended purposes.

Generalizability. During our last major experiences with performance assessment in the mid 1990s, one of the issues that derailed the movement toward rich, multisession, intentionally integrative performance tasks was between-task generalizability. We just could not demonstrate that a student's score would be consistent across these rich performance tasks (Linn, DeStefano, Burton, & Hansen, 1995; Pearson, DeStefano, & García, 1998; Shavelson, Baxter, & Pine, 1992). There are two distinct ways to interpret this problem. From a classic measurement perspective, if we cannot show inter-task stability, then we do not have a stable estimate of the construct that the tasks are designed to measure. But examined through the lens of the Text-Task Scenario approach, or even the more hermeneutic approach of some measurement theorists (e.g., Moss, 1996), it could be that the lack of generalizability across tasks reflects the reality of the situation. It could be that performance is not stable across certain features of different contexts (or, to use our language, across different Text-Task Scenarios). Something about the two different tasks—topical knowledge, interest, reasoning demands, or the like—renders them incomparable. This, of course, raises a genuine challenge for performance assessment in general and our Text-Task Scenario approach in particular: to what degree should we expect between-task stability across scenarios in our estimates of student achievement of any given construct under consideration?

If there are genuine limits on generalizability, then we need to know that. However, the problem also has an alternative conceptualization. We could ask the question, "Under what conditions (e.g., familiar content, high interest, transparent purpose) can a given student demonstrate construct competence—and under what other conditions might he or she be judged inadequate?" This is the kernel idea in Lipson and Wixson's (1986) notion of an interactive model of reading disability—that disability was not inherently situated within an individual but rather in the interaction of the very sorts of variables that define our Text-Task Scenarios. We are not sure how we, or the field, should come down on this generalizability dilemma, but we do know that our Text-Task Scenario approach brings the dilemma into full relief, for all of us to contemplate.

As an aside, we would note that the interactive nature of reading comprehension and implications for assessment suggest that Text-Task Scenarios are likely to function like individual performance tasks, because each presents readers with different texts, purposes for reading, comprehension tasks, and levels of student knowledge and interest. From a perspective that privileges difference over commonality, this is no different than the situation presented by current comprehension assessments, even those tests that employ short snippets of text with multiple-choice questions. It is just that when the data for standardized tests are examined, commonality trumps difference. So variations of the sort we consider consequential in our Text-Task Scenarios are not examined as systematic sources of variation that ought to be a part of our decision-making framework. Instead, they are considered to be noise, nuisance variation that gets relegated to the error term and renders our assessments slightly less reliable. They are not currently considered as a part of understanding performance. Because we believe that it is essential to consider how readers flexibly adapt and perform in the context of this complexity, we will need to conduct generalizability studies of Text-Task Scenarios as representative of the domain of reading comprehension, as well as through-course/interim assessments. But we do not begin these analyses with the assumption that performance will

necessarily generalize across Text-Task Scenarios. Instead, we need to understand which aspects of performance can be aggregated over tasks and time to represent stable aspects of performance, and conversely which aspects ought to be construed as unique sources of variation, either for groups or individuals, that ought to be fodder for specially tailored instructional intervention and differentiation.

Concluding Thoughts

We have argued that progress notwithstanding, we need to exert a lot more conceptual, curricular, and psychometric muscle to get reading comprehension assessment right. We need new models of assessment development that better reflect what we know to be true about reading comprehension—that it (a) is multifaceted; (b) varies in response to a range of skill, cognitive, motivational, procedural, disciplinary, contextual, and pedagogical forces; and (c) is amenable to instructional interventions that themselves deserve to be evaluated by valid, first-rate summative assessments and guided by equally stellar formative assessments. To achieve this goal, we believe that the field needs a new reading assessment R&D model, and the model we propose is to use Text-Task Scenarios to ensure that we examine reading comprehension in all of its possibilities and realities. We don't see any way to do justice to the Common Core State Standards for English language arts with a model that does any less. We hope that both PARCC and SBAC decide to accept this challenge and opportunity to make a difference in reading comprehension assessment, curriculum and pedagogy.

References

Alexander, P. A. (2003). Profiling the developing reader: The interplay of knowledge, interest, and strategic processing. In C. M. Fairbanks, J. Worthy, B. Maloch, J. V. Hoffman, & D. L. Schallert (Eds.), *The 52nd yearbook of the National Reading Conference* (pp. 47–65). Oak Creek, WI: National Reading Conference.

Alexander, P. A. (2005). The path to competence: A lifespan developmental perspective on Reading. *Journal of Literacy Research, 37*(4), 413–436.

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

Almond, R. G., & Mislevy, R. J. (1998). *Graphical models and computerized adaptive testing*. (TOEFL Tech. Rep. No. 14). Princeton, NJ: ETS.

Benjamin, R., Chun, M., Hardison, C., Hong, E., Jackson, C., Kugelmass, H., Nemath, A. et al. (2009). *Returning to learning in an age of assessment: Introducing the rationale of the Collegiate Learning Assessment* . Retrieved: December 16, 2011, from http://www.collegiatelearningassessment.org/

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74.

Bransford, J. D., Brown, A. L., & Cocking, R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

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.

Callahan, M., Benson-Griffo, V., & Pearson, P. D. (2009). Teacher knowledge and teaching reading. In F. Falk-Ross, S. Szabo, M. B. Sampson, & M. M. Foote (Eds.), *Literacy issues*

during changing times: A call to action, 30th yearbook of the College Reading Association (pp. 37-62). Logan, UT: College Reading Association.

Coh-Metrix (n.d.). Retrieved January 12, 2011 from http://cohmetrix.memphis.edu/cohmetrixpr/index.html

Coleman, D. (2010, December). *Implementing the Common Core Standards*. Presentation at the Association of Educational Publishers 2010 CEO Roundtable, New York, NY.

Darling-Hammond, L., & Pecheone, R. (2010, March). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Presented at the National Conference on Next Generation K–12 Assessment Systems, Center for K–12 Assessment & Performance Management with the Education Commission of the States (ECS) and the Council of Great City Schools (CGCS), Washington, DC.

Davis, F. B. (1944). Fundamental factors in reading comprehension. *Psychometrika, 9*, 185–197.

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

Destefano, L., Pearson, P. D., & Afflerbach, P. (1997). Content validation of the 1994 NAEP in reading: Assessing the relationship between the 1994 assessment and the reading framework. In R. Linn, R. Glaser, & G. Bohrnstedt (Eds.), *Assessment in transition: 1994 trial state assessment report on reading: background studies* (pp. 1–50). *Stanford, CA: The National Academy of Education*.

Doorey, N. (2011, February). Finding solutions, moving forward. In *Coming together to raise achievement: New assessments for the Common Core State Standards* (pp. 15–17). Princeton, NJ: Center for K–12 Assessment & Performance Management at ETS.

Johnson, D. D., & Pearson, P. D. (1975). Skills management systems: A critique. *The Reading Teacher, 28,* 757–764.

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: Longman.

Kirsch, I. S. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (ETS Research Rep. No. RR-01-25). Princeton, NJ: ETS.

Kirsch, I. S. (2003). Measuring literacy in IALS: A construct-centered approach. *International Journal of Educational Research, 39*, 181–190.

Lexile Framework for Reading. (n.d.). Retrieved January 12, 2011, from http://www.lexile.com/

Linn, R., DeStefano, L., Burton, E, & Hanson, M. (1995). Generalizability of New Standards Project 1993 Pilot Study Tasks in Mathematics. *Applied Measurement in Education, 9*, (2), 33-45. Lipson, M. Y., & Wixson, K. K. (1986). Research on reading disabilities: An interactionist perspective. *Review of Educational Research*, *56*, 111–136.

Mehan, H. (1993). Beneath the skin and between the ears. In S. Chaiklin & J. Lave (Eds.), *Understanding practice* (pp. 241–269). New York, NY: Cambridge University Press.

Moss, P. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher*, *25*, (1), 20-28. National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standardsfor English language arts & literacy in history/social studies, science, & technical subjects.* Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

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 scientificresearch 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.

Otto, W. (1977). The Wisconsin design: A reading program for individually guided elementary education. In R. A. Klausmeier, R. A. Rossmiller, & M. Saily (Eds.), Individually guided elementary education: Concepts and practices (pp. 216-237) New York, NY: Academic Press.

Otto, W. R., & Chester, R. D. (1976). *Objective-based reading*. Reading, MA: Addison-Wesley.

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems Competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Pearson, P. D., & DeStefano, L. (1993). Content validation of the 1992 NAEP in reading: Classifying items according to the reading framework. In R. Linn, R. Glaser, & G. Bohrnstedt (Eds.), *The trial state assessment: Prospects and realities: background studies*. Stanford, CA: The National Academy of Education.

Pearson, P.D., DeStefano, L., & García, G.E. (1998) Ten dilemmas of performance assessment. In C. Harrison and T. Salinger (Eds.) *Assessing reading 1: Theory and practice* (pp. 21-49). London: Routledge.

Pearson, P. D., & Hamm, D.N. (2005). The assessment of reading comprehension: A review of practices: past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Shavelson, R. J.; Baxter, G. P.; Pine, J. (1992). "Performance Assessments: Political Rhetoric and Measurement Reality," *Educational Researcher*, *21*, No. 4, pp. 22-27.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–304). New York, NY: Lawrence Erlbaum Associates.

Shepard, L. A. (2009). Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice* 28(3), 32-37

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

SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program application for new grants: Comprehensive assessment systems CFDA Number: 84.395B*. Retrieved from: http://www.k12.wa.us/SMARTER/RTTTApplication.aspx

Valencia, S., & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher, 40*, 726–733.