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.

2

The Assessment of Reading Comprehension: A Review of Practices—Past, Present, and Future[1]

P. David Pearson
University of California, Berkeley

Diane N. Hamm
Michigan State University

The purpose of this chapter is to build an argument for a fresh line of inquiry into the assessment of reading comprehension. We intend to accomplish that goal by providing a rich and detailed historical account of reading comprehension, both as a theoretical phenomenon and an operational construct that lives and breathes in classrooms throughout America. We review both basic research, which deals with reading comprehension largely in its theoretical aspect, and applied research, which is much more concerned about how comprehension gets operationalized in classrooms, reading materials, and tests.

With a renewed professional interest in reading comprehension (e.g., Rand Study Group, 2001), it is an optimal time to undertake a new initiative in the area of reading comprehension assessment. For a host of reasons, many having to do with curricular politics, reading comprehension has been placed on a back burner for well over 15 years. It is time it returned to a central role in discussions of reading. To do so, it needs our rapt and collective attention at this particular

[1]The original version of this chapter was prepared for the RAND Corporation as a background paper for their report to the Office of Educational Research and Improvement on Reading Comprehension.

point in history. First, reading comprehension, both its instruction and its assessment, is arguably the most important outcome of reform movements designed to improve reading curriculum and instruction. Second, given the national thirst for accountability, we must have better (i.e., conceptually and psychometrically more trustworthy) tools to drive the engines of accountability at the national, state, and local level. Third, and even more important, we need better assessments so that we can respond to the pleas of teachers desperate for useful tools to assist them in meeting individual needs. It is doubly appropriate that the assessment of reading comprehension receive as much attention as the construct itself. In the final analysis, a construct is judged as much by how it is operationalized as by how it is conceptualized.

The process of text comprehension has always provoked exasperated but nonetheless enthusiastic inquiry within the research community. Comprehension, or "understanding," by its very nature, is a phenomenon that can only be assessed, examined, or observed indirectly (Johnston, 1984a; Pearson & Johnson, 1978). We talk about the "click" of comprehension that propels a reader through a text, yet we never see it directly. We can only rely on indirect symptoms and artifacts of its occurrence. People tell us that they understood, or were puzzled by, or enjoyed, or were upset by, a text. Or, more commonly, we quiz them on "the text" in some way—requiring them to recall its gist or its major details, asking specific questions about its content and purpose, or insisting on an interpretation and critique of its message. All of these tasks, however challenging or engaging they might be, are little more than the residue of the comprehension process itself. Like it or not, it is precisely this residue that scholars of comprehension and comprehension assessment must work with to improve our understanding of the construct. We see little more of comprehension than Plato saw of the shadows in the cave of reality.

Models of reading comprehension and how to assess it have evolved throughout the 20th century (see Johnston, 1984b). Many techniques of assessment have risen to prominence and then fallen out of use, some to be reincarnated decades later, usually with new twists. Our aim is to provide a thorough account of what we know about assessing reading comprehension. Where possible and appropriate, we take detours into research and theory about the comprehension process, on the grounds that conceptions of the process, because they have influenced how it is assessed, will inform our understanding. We hope to illuminate the patterns, cycles, and trends in comprehension assessment. Through these efforts, we hope to provide our readers with a means to evaluate the current state of reading assessment, which we believe has reached a critical juncture, one that can be crossed only by shaping a research agenda that will improve our capacity to create valid, fair, and informative assessments of this important phenomenon.

HISTORICAL FOUNDATIONS OF READING COMPREHENSION ASSESSMENT

Before the Beginning

Although reading comprehension assessment as a formal, identifiable activity is a 20th century phenomenon, it has been a part of classrooms as long as there have been schools, required texts, students who are required to read them, and teachers wanting or needing to know whether students understood them. In every century and every decade, every assignment given by a teacher, every book report or chapter summary, and every conversation about a book, story, article, or chapter, has provided an opportunity for assessment. It was not until well into the 20th century that we began to seize those opportunities. There are two plausible explanations for the relatively late arrival of comprehension as an indicator of reading achievement. First, the default indicator of reading prowess in the 17th to 19th centuries was definitely oral capacity, indexed either by accuracy or by expressive fluency, in the tradition of declamation and oratory (see Mathews, 1996, or Smith, 1986, for accounts of this emphasis). Second, within ecclesiastical circles, comprehension, at least in the sense of personal understanding, was not truly valued; if it mattered, it mattered largely as a stepping stone to the more valued commodity of text memorization (see Olson, 1994, for an account of the various religious traditions in text interpretation).

The Beginning

It is well worth our effort to examine early trends in reading assessment, for they suggest that nearly all of the tools we use to measure reading comprehension today made an appearance in some way shape or form before World War II. Granted, today's formats and approaches may look more sophisticated and complex, but, as our review demonstrates, those formats were there, at least in prototypic form, long ago.

The first systematic attempts to index reading ability by measuring comprehension date back to the period just prior to World War I. Binet, as early as 1895 (cited in Johnston, 1984a), used comprehension test items, ironically, to measure intelligence rather than reading achievement. In 1916, Kelly brought us the first published comprehension assessment, the *Kansas Silent Reading Tests*. Thorndike, in his classic 1917 piece, *Reading as Reasoning: A Study of Mistakes in Paragraph Reading*, offered us our first professional glimpse "inside the head" as he tried to characterize what must have been going on in the minds of students to produce the sorts of answers they come up with when answering questions about text. As we indicated earlier, the quest to get as close as possible to the

"phenomenological act of comprehension" as it occurs has always driven researchers to discover new and more direct indexes of reading comprehension.

The scientific movement and the changing demographic patterns of schooling in America were both forces that shaped the way reading was conceptualized and assessed in the first third of the century. Schools had to accommodate rapid increases in enrollment due to waves of immigration, a rapidly industrializing society, the prohibition of child labor, and mandatory school attendance laws. The spike in school enrollment, coupled with a population of students with dubious literacy skills, dramatically increased the need for a cheap, efficient screening device to determine students' levels of literacy. During this same period, psychology struggled to gain the status of a "science" by employing the methods that governed physical sciences and research. In America, the behaviorist schools of thought, with their focus on measurable outcomes, strongly influenced the field of psychology (Johnston, 1984a; Pearson, 2000; Resnick, 1982); quantification and objectivity were the two hallmarks to which educational "science" aspired. Thus, when psychologists with their newfound scientific lenses were put to work creating cheap and efficient tests for beleaguered schools, the course of reading assessment was set. Group administered, multiple-choice, standardized tests would be the inevitable result.

The other strong influence in moving toward comprehension as a measure of reading accomplishment was the curricular shift from oral to silent reading as the dominant mode of reading activity in our classrooms. Although the first published reading assessment, circa 1914 (Gray, 1916, 1917), was an oral reading assessment created by William S. Gray (who eventually became a preeminent scholar in the reading field and the senior author of the country's most widely used reading series), most reading assessments developed in the first third of this century focused on the relatively new construct of silent reading (see Pearson, 2000; Johnston, 1984a). Unlike oral reading, which had to be tested individually and required that teachers judge the quality of responses, silent reading comprehension (and rate) could be tested in group settings and scored without recourse to professional judgment; only stop watches and multiple-choice questions were needed. In modern parlance, we would say that they moved from a "high inference" assessment tool (oral reading and retelling) to a "low inference" tool (multiple-choice tests or timed readings). Thus, it fit the demands for efficiency and scientific objectivity, themes that were part of the emerging scientism of the period. The practice proved remarkably persistent for at least another 40 or 50 years. Significant developments in reading comprehension would occur in the second third of the century, but assessment would remain a psychometric rather than a cognitive activity until the cognitive revolution of the early 1970s.

It is important to note that comprehension instruction and the curricular materials teachers employed were driven by the same infrastructure of tasks used to create test items—finding main ideas, noting important details, determining sequence of events, cause–effect relations, comparing and contrasting, and drawing conclusions.[2] If these new assessments had not found a comfortable match in school curricular schemes, one wonders whether they would have survived and prospered to the degree that they did.

Intelligence and Comprehension. Interestingly, it was difficult to tell the difference, in these early years, between reading comprehension assessments and intelligence tests. Freeman (1926) noted that Binet (1895, as cited in Johnston, 1984a) had used reading comprehension items as a part of his IQ battery. Consider, also, this item from an early (but undated) edition of a Thurstone (n.d.) intelligence test[3] (cited in Johnson, 1984a):

> Every one of us, whatever our speculative opinion, knows better than he practices, and recognizes a better law than he obeys. (Froude)

Check two of the following statements with the same meaning as the quotation above.

_____ To know right is to do the right.
_____ Our speculative opinions determine our actions.
_____ Our deeds often fall short of the actions we approve.
_____ Our ideas are in advance of our every day behavior.

Minor Anomalies and Omens of the Future. Although behaviorism was the dominant paradigm underlying curricular and assessment work during this period, remnants of a cognitively more complex approach of the sort that Huey described near the turn of the century (Huey, 1908) made minor appearances on the assessment scene. Free recall was used by a few researchers as an index of comprehension. Starch (1915), for example, created a ratio (the number of relevant words a student remembered in a passage in comparison to the proportion of total words remembered) as an index of comprehension. Courtis (1914) developed a similar but simpler index (ratio of idea units reproduced or interpreted to the number possible). These indexes, especially the relevance index,

[2]This tradition of isomorphism between the infrastructure of tests and curriculum has been a persistent issue throughout the century. See, for example, Johnson & Pearson (1975), and Resnick (1982). See also Smith (1966) for an account of the expansion of reading comprehension as a curricular phenomenon.

[3]The use of more than one right answer predates the infamous a, b, c (a and b) multiple-choice format as well as the systematic use of the "more than one right answer" approach used in some state assessments in the 1980s and 1990s (Pearson et al., 1990).

foreshadow work in the 1970s and 1980s on "importance" (as indexed by the relevance of propositions to a text's ideational structure (e.g., Rumelhart, 1977). Even at this early stage, scholars recognized that recall is not the same process as making or uncovering meaning (Kelly, 1916), but recall continued to be used in research, and later in practice, as a direct index of comprehension. This use of recall would be revived in the 1970s as a retelling procedure, which would give us a window on whether students were remembering important ideas in stories (Stein & Glenn, 1977) or in the propositional data base of expository texts (Kintsch & van Dijk, 1978; Turner & Greene, 1977).

Consistent with the efficiency criterion in the new scientific education, speed was often used as an important factor in assessing comprehension. Kelly, the author of the *Kansas Silent Reading Tests* (1916), required students to complete as many of a set of 16 diverse tasks as they could in the 5 min allotted. The tasks included some "fill in the blanks," some verbal logic problems, and some procedural tasks (following directions). Monroe also used a speeded task—asking students to underline the words that answered specific questions.

We can even find foreshadowing of the error detection paradigms that were to be so widely used by psychologists investigating metacognitive processes in the 1970s through the 1990s (Markman, 1977; Winograd & Johnston, 1980). For example, Chapman (1924) asked students to detect words that were erroneous or out of place in the second half of each paragraph (presumably they did so by using, as the criterion for rejection, the set or schema for paragraph meaning that became established as they read the first half). In 1936, Eurich required students to detect "irrelevant clauses" rather than words.

Thorndike (1917) was probably the first educational psychologist to try to launch inquiry into both the complex thought processes associated with comprehension and assessment methods. He referred to reading "as reasoning," suggesting that there are many factors that comprise it: "elements in a sentence, their organization … proper relations, selection of certain connotations and the rejection of others, and the cooperation of many forces" (Thorndike, 1917, p. 323). He proposed ideas about what should occur during "correct reading," claiming that a great many misreadings of questions and passages are produced because of underpotency or overpotency of individual words, thus violating his "correct weighting" principle: "Understanding a paragraph is like solving a problem in mathematics. It consists in selecting the right elements in the situation and putting them together in the right relations, and also with the right amount of weight or influence or force of each" (Thorndike, 1917, p. 329). Of course, he assumed that there are such things as "correct" readings. He argued further that in the act of reading, the mind must organize and analyze ideas from the text. "The vice of the poor reader is to say the words to himself without ac-

tively making judgments concerning what they reveal" (Thorndike, 1917, p. 332). Clearly for Thorndike, reading was an active and complex cognitive process. Although this perspective did not become dominant in this early period, it certainly anticipated the highly active view of the reader that would become prominent during the cognitive revolution of the 1970s.[4]

Paralleling an active line of inquiry in oral reading error analysis (see Allington, 1984) during this period, some researchers followed Thorndike's lead and tried to develop taxonomies of the kinds of errors readers make either during decoding or understanding. Touton and Berry (1931, cited in Davis, 1968) classified errors into six categories based on research on college students' (a) failure to understand the question, (b) failure to isolate elements of "an involved statement" read in context, (c) failure to associate related elements in a context, (d) failure to grasp and retain ideas essential to understanding concepts, (e) failure to see setting of the context as a whole, and (f) other irrelevant answers.

Although Goodman (1968, 1969) is rightfully credited with helping us understand that oral reading errors, or *miscues*, to use his term, can reveal much about the comprehension processes in which a student engages; there were inklings of this perspective emerging in the 1920s. Gates (1937), for example, was interested in how readers' fluency may be an indicator of one's ability and understanding. He looked at readers' "error of hesitation," that is, whether a reader stumbled over a word or phrase. Durrell (1955) and later Betts (1946) sought to use these error patterns as indicators of the level of reading material students could handle, both from a word recognition and comprehension perspective. These early scholars determined that students who misread many words (they found that 2% seems to be our outside limit—although modern scholars often go up to 5%) will have difficulty comprehending a passage. These harbingers notwithstanding, it would be another 30 years before Goodmans' (Goodman, 1968; Goodman, 1969; Goodman & Burke, 1970) miscue analysis work prompted us to take oral reading miscues seriously as a lens that would allow us to look into the windows of the mind at the comprehension process.

PSYCHOMETRICS GATHERS MOMENTUM

Two significant events in the history of assessment occurred during the 1930s and 1940s; both would have dramatic effects on reading comprehension assess-

[4]It is somewhat ironic that the sort of thinking exhibited in this piece did not become a dominant view in the early 1900s. Unquestionably, Thorndike (1917) was the preeminent educational psychologist of his time. Further, his work in the psychology of learning (the law of effect and the law of contiguity) became the basis of the behaviorism that dominated educational psychology and pedagogy during this period, and his work in assessment was highly influential in developing the components of classical measurement theory (reliability and validity). Somehow this more cognitively-oriented side of his work was less influential, at least in the period in which it was written.

ment. First, in 1935, IBM introduced the IBM 805 scanner, which had the potential to reduce the cost of scoring dramatically (compared to hand-scoring of multiple-choice, or "even worse," short answer and essay tests) by a factor of 10 (Johnston, 1984a). It is not insignificant that the Scholastic Aptitude Test, which, in the 1920s and early 1930s, had been mostly an essay test, was transformed into a machine-scorable multiple-choice test shortly thereafter (Resnick, 1982). This development paved the way for a new generation of multiple-choice assessments for all fields in which testing is used; reading comprehension assessment proved no exception.

Determining the Infrastructure of Reading Comprehension

The second important event was the publication, in 1944, of Frederick Davis's landmark doctoral dissertation in which he used a brand new statistical tool, factor analysis, to determine whether a set of conceptually distinct subtests of reading comprehension (entities like finding main ideas, selecting details, determining word meanings, drawing conclusions, determining cause–effect relations, distinguishing fact from opinion, and the like) were also psychometrically distinct. Factor analysis is a technique, still highly popular among traditional psychometricians, in which the covariation among "units" (usually items or subtests) is examined to discover which units tend to cluster with (i.e., covary with) which other units. Armed with this new tool, researchers were (at least theoretically) ready to answer a question that had vexed both test makers and curriculum designers for the two or three decades in which reading tests and reading curriculum had become part of the American educational landscape: Is comprehension a unitary or a multivariate construct? That is, are there distinct subcomponents, subprocesses, or "skills" that ought to be measured and perhaps taught separately? Or, alternatively, is reading better construed as a unitary process that ought to be considered holistically?

In his groundbreaking 1944 study, Davis reviewed the literature describing reading comprehension as a construct and found several hundred skills mentioned. He sorted them into nine categories (see Table 2.1) that he felt constituted conceptually distinct groups; from these he devised nine testable skills (based also in part on correlation data). Davis employed the most sophisticated factor analytic tools available (Kelley, 1935) in his search for psychometric uniqueness to match the conceptual uniqueness of his categories. Acknowledging the unreliability of some of the subtests (due among other factors to the small standard deviations and the fact that each passage had items from several cognitive categories attached to it), he was able to conclude that reading comprehension consisted of two major factors, word knowledge and "reasoning

TABLE 2.1
Davis's Nine Potential Factors

1. Word meanings	6. Text-based questions with paraphrase
2. Word meanings in context	7. Draw inferences about content
3. Follow passage organization	8. Literary devices
4. Main thought	9. Author's purpose
5. Answer specific text-based questions	

about reading," that were sufficiently powerful and reliable to guide us in the construction of tests and reading curriculum. He speculated that another three factors (comprehension of explicitly stated ideas, understanding passage organization, and detecting literary devices) had the potential, with better item development, to reveal themselves as independent factors.

Between 1944 and the early 1970s, several scholars attempted to either replicate or refute Davis's findings. Harris (1948) found only one factor among the seven he tested. Derrik (1953) found three, and they were consistent across different levels of passage length. Hunt (1957) used differential item analysis and correction formulae to adjust his correlations, finding vocabulary (i.e., word knowledge) as the single most important factor.

Partially in response to the conflicting evidence in the field, Davis (1968, 1972) replicated his earlier work but with a more sophisticated design and set of analysis tools, not to mention the newly available capacity of mainframe computers. Using a very large set of items from various publishers and his own bank of items, he constructed 40 multiple-choice questions per hypothesized skill, each derived from a separate passage.[5] From this set, he created a more psychometrically sound subset of 24 items for each skill; that is, he selected items that exhibited high correlations with the total subtest and the total test but low correlations with other subtests. Armed with this virtually optimal set of items, he tested the independence of eight distinguishable subskills remarkably similar to the set used in his 1944 study (see Table 2.2).

Davis (1968) used cross validation (the use of multiple regression weights computed in one matrix to obtain multiple correlation coefficients in a different but analogous matrix) and multiple regression analyses to determine the pro-

[5]The availability of items, each derived from a separate passage, represents a great psychometric advantage because the conceptual case for item independence can be made. However, as a practical matter, it is dubious that we would, could, or will ever use comprehension assessments with only one item per passage.

TABLE 2.2
Davis's Eight Potential Factors (1968)

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

portion of common and unique variance among the subtests. Remembering word meanings explained the most (32%) unique variance. This was followed by "drawing inferences from content," with 20% unique variance. This was followed, in order of magnitude of unique variance, by "structure of the passage," "writerly techniques" and "explicit comprehension." Again, he concluded that comprehension was not a unitary factor.

During the late 1960s and early 1970s, there was a flurry of activity in this tradition. Spearitt (1972) reviewed and re-analyzed what seemed to be Davis's 1968 work, finding at least four statistically differentiable skills, only one of which appeared to be clearly unique, whereas the other three could be measuring the same general ability. Schreiner, Hieronymus, and Forsyth (1971), analyzing data from the popular Iowa Test of Basic Skills, supplemented by additional subtests measuring general cognitive and verbal skills (e.g., listening comprehension, verbal reasoning, and reading and skimming speed), found that all of the supplementary tests were statistically independent but that the various reading comprehension subtests (paragraph meaning, cause and effect, reading for inferences, and selecting main ideas) could not be statistically differentiated from one another. 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 would not, however, stop the proliferation of single skill tests, which became even more popular in the 1970s and 1980s.

Although it is difficult to trace precise causal links between this extensive work on factor analysis and the prevailing practices in reading assessment, it is worth noting that virtually all of the commercially popular reading assessments of this era (for that matter, in both preceding and following eras)[6] followed the

[6]The persistence and resilience of standardized reading tests is quite remarkable. As a part of our work, we traced the evolution, across several editions, of several of the most popular standardized tests. Stability clearly outshines change in the examination of the history of these tests.

practice of embedding comprehension skills that varied on some dimension of perceived cognitive complexity (finding details, inferring details, selecting or inferring main ideas, drawing inferences about characters and ideas, detecting author's craft, etc.) within a set of short passages on different topics.[7] To achieve a balanced test, developers would build a matrix in which content (in the case of reading comprehension, content is construed as the topics of the passages) was crossed with the processes (the various cognitive skills in the underlying model of reading); they would then use the matrix to monitor the "balance" of item types appearing in the final version of a commercial test.

The Cloze Procedure

In the 1950s, Wilson Taylor (1953) developed the cloze procedure as an alternative to the conventional standardized test. Taylor began with the assumption that even the process of writing multiple-choice items was subjective. Instead of introducing subjectivity by requiring test developers to determine what content and features of a passage should be assessed, Taylor developed the cloze technique, which replaces human judgment with a mechanical approach to item development. A test designer simply deletes every nth word (usually every fifth word) in a passage; the task of the examinee is to fill in each cloze blank. The more blanks filled in, the higher the comprehension score. There was a buzz of excitement about the cloze procedure during the 1960s and 1970s (see Rankin, 1965; see also Bormuth, 1966, for the most elaborate application of cloze). Cloze was touted as the scientific alternative to multiple-choice tests of reading comprehension. It was widely used as the comprehension criterion in studies of readability in the 1960s (see Bormuth, 1966). It became the cornerstone of reading assessment for ESL (English as a second language) speakers (see Bachman, 1982), where it is still widely used (Bachman, 2000).

Cloze has experienced a great deal of adaptation over the years. For example, in the classic cloze procedure, students are asked to write in their responses when every fifth word is deleted. Only exact replacement is scored as correct; synonyms will not do. However, researchers and test developers have created a modified cloze procedure using a whole host of variations:

- Allow synonyms to serve as correct answers.
- Delete only every fifth content word (leaving function words intact).
- Use an alternative to every fifth word deletion.

[7]Highly influential in this period was Bloom's (1956) taxonomy of the cognitive domain, in which he laid out a hierarchy of processes that presumably varied on a dimension of cognitive complexity.

- Delete words at the end of sentences and provide a set of choices from which examinees are to pick the best answer (this tack is employed in several standardized tests, including the Stanford Diagnostic Reading Test and the Degrees of Reading Power).

The unsettled question about cloze tests is whether they are measures of individual differences in comprehension or measures of the linguistic predictability of the passages to which they are applied. They have been widely criticized for this ambiguity. But perhaps the most damaging evidence in their role as indexes of reading comprehension is that they are not sensitive to "intersentential" comprehension, that is, understanding that reaches across sentences in a passage. In a classic study, Shanahan, Kamil, and Tobin (1983) created several passage variations and assessed cloze fill-in rates. In one condition, sentence order was scrambled by randomly ordering the sentences. In another condition, sentences from different passages were intermingled, and in a third condition, isolated sentences from different passages were used. There were no differences in cloze fill-in rate across any of these conditions, indicating that an individual's ability to fill in cloze blanks does not depend on passage context; in short, when people fill in cloze blanks, they do not think across sentence boundaries. In the period of the cognitive revolution of the 1980s, in which comprehension was viewed as an integrative process, a measure that did not require text integration did not fare well.

These findings notwithstanding, modified, multiple-choice versions of cloze are still alive and well in standardized tests (i.e., the Degrees of Reading Power and the Stanford Diagnostic Reading Test referred to earlier) and in ESL assessment for adults and college students (Bachman, 2000).

Passage Dependency

Beginning in the late 1960s, a new construct arose in reading assessment, one that, at the time, had the impact of a "the emperor has no clothes" epiphany. Several scholars became concerned about the fact that many of the questions of reading comprehension on standardized tests could be answered correctly without reading the passage (mainly because the information assessed was likely to exist in examinees' prior knowledge as well as in the text). This problem is particularly exacerbated in passages about everyday or common academic topics (in comparison, for example, to fictional narratives). A number of researchers (e.g., Tuinman, 1974, 1978) conducted passage dependency studies in which some participants took the test without the passage being present. The difference between the *p* value of an item in the two conditions (with and without

text) is an index of an item's passage dependency. The logic of this construct is simple and compelling: a reader should have to read a passage to answer questions about it. Interestingly, the interest in passage dependency, like the interest in cloze, waned considerably during the cognitive revolution. In the new paradigm, prior knowledge would be embraced as one of the cornerstones of comprehension, and scholars would attempt to take prior knowledge into account rather than try to eliminate or encapsulate its impact on comprehension (see Johnston, 1984b, for an account of these attempts during the early 1980s).

THE IMPACT OF CRITERION-REFERENCED ASSESSMENT[8]

Beginning in the 1960s and lasting until the late 1980s, criterion-referenced tests (CRT) became a major force in classroom reading assessment and basal reading programs. The theoretical rationale for CRTs comes from the mastery learning work of eminent scholars such as Benjamin Bloom (1968), John Carroll (1963), and Robert Gagné. The idea behind mastery learning was that if we could just be more precise about the essential elements involved in learning any particular domain or process, we could bring most, if not all, students to higher levels of achievement, perhaps even levels where we could state with confidence that they had mastered the skill. The precision could be achieved, according to the champions of mastery learning (see Otto, 1977; Otto & Chester, 1976), by decomposing the domain or process into essential elements. Then one could teach (and test) each of the elements to mastery. To determine whether an element had been mastered, a transparent test was needed, one in which a clear criterion had been met. In CRTs, an absolute standard, such as 80% correct on a test of the particular element, not a relative standard, such as the average score of students in a given classroom, grand, school, or state, would serve as the criterion.[9]

Criterion-referenced assessments were popular throughout our schools and curricula in the 1970s and 1980s, but nowhere was their influence more dramatically felt than in basal reading materials. Starting in the early 1970s, basal programs developed criterion-referenced tests for every unit (a grouping of 6 to 8 stories plus associated activities) and every book in the series. Each successive edition of basal programs brought an increase in the number of these single component tests—tests for each phonics skill (all the beginning, middle and fi-

[8]For this section, we relied heavily on the treatment of these issues in Sarroub and Pearson (1998).

[9]The most compelling version of criterion-referenced assessment is domain-referenced assessment, a practice in which the items in a test are viewed as a sample from the population of items representing performance in that domain of knowledge or inquiry. Early proponents of domain referenced assessment (see Bormuth, 1970; Hively, 1974) saw great hope for this approach as a way of estimating student mastery over knowledge domains. Modern proponents (e.g., Bock, 1997) are no less enthusiastic.

nal consonant sounds, vowel patterns, and syllabication), tests for each comprehension skill (main idea, finding details, drawing conclusions, and determining cause–effect relations) at every grade level, tests for alphabetical order and using the encyclopedia, and just about any other skill one might imagine. With powerful evidence from mastery learning's application to college students (Bloom, 1968), publishers of basal programs and some niche publishers began to create and implement what came to be called skills management systems.[10] In their most meticulous application, these systems became the reading program. Students took a battery of mastery tests, practiced those skills they had not mastered (usually by completing worksheets that looked remarkably like the tests), took tests again, and continued through this cycle until they had mastered all the skills assigned to the grade level (or until the year ended). Unsurprisingly, the inclusion of these highly specific skill tests had the effect of increasing the salience of workbooks, worksheets, and other "skill materials" that students could practice in anticipation of (and as a consequence of) mastery tests. Thus, the basals of this period included two parallel systems: (a) a graded series of anthologies filled with stories and short nonfiction pieces for oral and silent reading and discussion, and (b) an embedded skills management system to guide the development of phonics, comprehension, vocabulary, and study skills. In the true mastery programs (e.g., Board of Education, City of Chicago, 1984; Otto, 1977) and in some basal programs, students who failed a particular subtest were required to practice skill sheets that looked remarkably like the mastery tests until they could achieve mastery (which was usually and arbitrarily defined as 80% correct).

For comprehension assessment, the consequences were dramatic. Even with standardized, multiple-choice assessments of the ilk studied by Davis (1968), there had been some sense that important aspects of a passage ought to be queried. But with the new criterion-referenced assessments of reading comprehension, the number of comprehension subskills increased dramatically, as did the number of specific skill tests for each of these. Clearly, in these assessments, as illustrated in Fig. 2.1, the emphasis is on the skill rather than the passage. The passage is nothing more than a vehicle that allows for an assessment of the skill. For example, Figure 1 tests a child's ability to recognize sequential order.

[10]The most popular of these systems was the Wisconsin Design for Reading Skill Development, followed closely by Fountain Valley. Their heyday was the decade of the 1970s, although they remained a staple, as an option, through the 1980s and 1990s and are still available as options in today's basals. For an account of the rationale behind these systems, see Otto (1977). For a critique of these programs during their ascendency, see Johnson and Pearson (1975).

The children wanted to make a book for their teacher. One girl brought a camera to school. She took a picture of each person in the class. Then they wrote their names under the pictures. One boy tied all the pages together. Then the children gave the book to their teacher.

1. What happened first?
   a. The children wrote their names
   b. Someone brought a camera to school
   c. The children gave a book to their teacher
2. What happened after the children wrote their names?
   a. A boy put the pages together
   b. The children taped their pictures
   c. A girl took pictures of each person
3. What happened last?
   a. The children wrote their names under the pictures
   b. A girl took pictures of everyone
   c. The children gave a book to their teacher

FIG. 2.1. An example of a basal reader's criterion-referenced test (adapted from the Ginn Reading Program, 1982).

The basals of the 1970s and 1980s were filled with tests like these for up to 30 different comprehension skills (Johnson & Pearson, 1975).[11] More importantly, they persisted and flourished in the face of many professional critiques of their theoretical and practical efficacy, validity, and utility (Johnson & Pearson, 1975; Valencia & Pearson, 1987b).

Standardized reading assessments also felt the impact of criterion-referenced assessment. First, several testing companies brought out their own versions of criterion-referenced assessments that could compete with those offered by start-up companies (e.g., Fountain Valley and National Computer Systems —the marketer of Wisconsin Design). Second, and perhaps more influential, most testing companies created the capacity to report scores on their standardized assessment by specific comprehension subskills (e.g., main idea, details, inferences, author's craft, etc.). In so doing, they were, in somewhat ironic a fashion, completing the goal laid out by Davis (1944) when he began his quest

[11]A related movement, domain-referenced assessment, also became popular during this period, but it did not establish a strong foothold within reading assessment. In a field like mathematics, it made sense to talk about the domain or population of all two-digit multiplication facts or the like. However, a concept such as all of the possible literal comprehension probes for stories appropriate for Grade 2 readers seems to make little sense. Only Bormuth (1970) developed anything approaching a domain-referenced approach for assessing reading comprehension in his nearly forgotten classic, *On the Theory of Achievement Test Items*.

to identify independent subprocesses of comprehension in the early 1940s. Thus, by 1985, it was possible for a school to receive not only conventional reports of the average, by grade or class, of grade norm, percentile, NCE (normal curve equivalent), or stanine scores, but also reports of the percentage of students, by grade or class, who demonstrated mastery (i.e., exceeded the cut score) on each of the component skills included in the test.

THE REVOLUTIONS BEGIN

Somewhere during this period of active psychometric work on reading assessment (between 1955 and 1975—the exact point of departure is hard to fix), the field of reading experienced a paradigm shift. The underlying theoretical perspectives of behaviorism on which reading pedagogy and reading assessment had been built throughout the century were challenged, overturned, and replaced by a succession of pretenders to the theoretical throne. Along the way, reading scholars were forced to confront fundamental shifts in the prevailing views of reading and writing that led to the creation of a variety of serious curricular alternatives to the conventional wisdom of the 1970s. Reading became an ecumenical scholarly commodity; it was embraced by scholars from many different fields of inquiry (see Pearson & Stephens, 1993, for a complete account of this phenomenon). The first to take reading under their wing were the linguists, who wanted to convince us that reading was a language process closely allied to its sibling language processes of writing, speaking, and listening. Then came the psycholinguists and the cognitive psychologists, who convinced us to seek out the underlying language and cognitive processes that enabled reading. They were soon followed soon by the sociolinguists, the philosophers, the literary critics, and the critical theorists, each bringing a critique of its immediate predecessor, each offering a new perspective to guide instructional practice, including assessment.

It is not altogether clear why reading has attracted such interest from scholars in so many other fields. One explanation is that reading is considered by so many to be a key to success in other endeavors in and out of school; this is often revealed in comments like, "Well if you don't learn to read, you can't learn other things for yourself." Another reason is that scholars in these other disciplines thought that the educationists had got it all wrong: it was time for another field to have its perspective heard. Whatever the reasons, the influence of these other scholarly traditions on reading pedagogy is significant; in fact, neither the pedagogy nor the assessment of the 1980s and 1990s can be understood without a firm grounding in the changes in worldview that these perspectives spawned.

In terms of reading comprehension assessment, three of these movements are particularly important: cognitive psychology, sociolinguistics (and more general sociocultural perspectives), and literary theory (in the form of reader re-

sponse theory). Cognitive psychology spawned the first of two major shifts in comprehension assessment; the second was prompted by the joint influence sociolinguistics and literary theory.

COGNITIVE PSYCHOLOGY

In rejecting behaviorism, cognitive psychology allowed psychologists to extend constructs such as human purpose, intention, and motivation to a greater range of psychological phenomena, including perception, attention, comprehension, learning, memory, and executive control or "metacognition" of all cognitive process. All of these would have important consequences in reading pedagogy and, to a lesser extent, reading assessment.

The most notable change within psychology was that it became fashionable for psychologists, for the first time since the early part of the century, to study complex phenomena such as language and reading.[12] And in the decade of the 1970s, works by psychologists flooded the literature on basic processes in reading. One group focused on text comprehension, trying to explain how readers come to understand the underlying structure of texts. We were offered story grammars—structural accounts of the nature of narratives, complete with predictions about how those structures impede and enhance story understanding and memory (Mandler & Johnson, 1977; Rumelhart, 1977; Stein & Glenn, 1977). Others chose to focus on the expository tradition in text (e.g., Kintsch, 1974; Meyer, 1975). Like their colleagues interested in story comprehension, they believed that structural accounts of the nature of expository (informational) texts would provide valid and useful models for human text comprehension. And in a sense, both of these efforts worked. Story grammars did provide explanations for story comprehension. Analyses of the structural relations among ideas in an informational piece also provided explanations for expository text comprehension. However, neither text-analysis tradition really tackled the relation between the knowledge of the world that readers bring to text and comprehension of those texts. In other words, by focusing on structural rather than the ideational, or content, characteristics of texts, they failed to get to the heart of comprehension. That task, as it turned out, fell to one of the most popular and influential movements of the 1970s: schema theory.

Schema theory (see Anderson & Pearson, 1984; Rumelhart, 1981) is a theory about the structure of human knowledge as it is represented in memory. In our memory, schemata are like little containers into which we deposit the par-

[12]During this period, great homage was paid to intellectual ancestors such as Edmund Burke Huey, who, as early as 1908, recognized the cognitive complexity of reading. Voices such as Huey's, unfortunately, were not heard during the period from 1915 to 1965 when behaviorism dominated psychology and education.

ticular traces of particular experiences as well as the "ideas" that derive from those experiences. So, if we see a chair, we store that visual experience in our "chair schema." If we go to a restaurant, we store that experience in our "restaurant schema." If we attend a party, we store that experience in our "party schema," and so on.

Schema theory also provided a credible account of reading comprehension, which probably, more than any of its other features, accounted for its popularity within the reading field in the 1970s and 80s.[13] Schema theory struck a sympathetic note with researchers as well as practitioners. It provided a rich and detailed theoretical account of the everyday intuition that we understand and learn what is new in terms of what we already know. It also accounted for the everyday phenomenon of disagreements in interpreting stories, movies, and news events—we disagree with one another because we approach the phenomenon with very different background experiences and knowledge.

With respect to reading comprehension, schema theory encouraged educators to examine texts from the perspective of the knowledge and cultural backgrounds of students to evaluate the likely connections that they would be able to make between ideas inscribed[14] in the text and the schema that they would bring to the reading task. Schema theory also promoted a constructivist view of comprehension; all readers, at every moment in the reading process, construct the most coherent model of meaning for the texts they read.[15] Perhaps the most important legacy of this constructivist perspective was that it introduced ambiguity about the question of where meaning resides. Does it reside in the text? In the author's mind as she sets pen to paper? In the mind of each reader as she builds a model of meaning unique to her experience and reading? In the interaction between reader and text? Schema theory raised, but did not settle these questions.

The Impact of Cognitive Science on Assessment

The impact of this new work in comprehension on curriculum and classroom teaching was immediate. We saw huge changes in basal readers, which, even

[13]It is not altogether clear that schema theory is dead, especially in contexts of practice. Its role in psychological theory is undoubtedly diminished due to attacks on its efficacy as a model of memory and cognition. See McNamara, Miller, and Bransford (1991) or Spiro and Jehng (1990).

[14]Smagorinsky (2001) used the phrase *inscribed* in the text as a way of indicating that the author of the text had some specific intentions when he or she set pen to paper, thereby avoiding the thorny question of whether meaning exists "out there" outside of the minds of readers. We use the term here to avoid the very same question.

[15]Most coherent model is defined as that model which provides the best account of the "facts" of the text uncovered at a given point in time by the reader in relation to the schemata instantiated at that same point in time.

until the late 1980s, remained the core tool of classroom practice. These included the following: (a) more attention to the role of prior knowledge introducing new texts, explicit teaching of comprehension strategies; (b) attention to text structure (in the form of story maps and visual displays to capture the organizational structure of text); and (c) the introduction of metacognitive monitoring (reflecting on what one has read, said, or written to see if it makes sense; see Pearson, 2000).

The impact on assessment, in particular, the unsettled question of where meaning resides, was fairly transparent: How, with even a modicum of respect for fairness, can we use tests with single correct answers if we know that answers are influenced by experience and background knowledge? It was not long before educators began to ask questions about whether the long tradition of standardized, multiple-choice assessments could or should continue to be used as measures of program quality or teacher effectiveness.

Table 2.3 (from Valencia & Pearson, 1987b) illustrates clearly the tensions that existed between the new cognitively oriented views of the reading process and prevailing assessment praxis in the mid 1980s.

By the late 1980s, constructivist approaches to reading assessment began to emerge. These were new efforts and new perspectives, and they sought new formats and new approaches to question generation for assessments. They privileged conceptual over psychometric criteria in building new reading assessments. They emphasized the need for assessments to reflect resources such as prior knowledge, environmental clues, the text itself, and the key players involved in the reading process. They emphasized metacogntion as a reflective face of comprehension. And they championed the position that only a fresh start in assessments would give us tests to match our models of instruction.

Major Changes. Changes included longer text passages, more challenging questions, and different question formats (such as the "more than one right answer" format and open-ended questions). Reading scholars acknowledged that although all multiple-choice items include answers that are plausible under certain conditions, they do not necessarily invite reflection or interactive learning. Assessment efforts in Illinois and Michigan (see Valencia, Pearson, Peters, & Wixson, 1989) led the charge in trying to incorporate these new elements. First, in the spirit of authenticity, they included longer, and more naturally occurring or "authentic" text selections in tests. Also, both included test items that measured prior knowledge rather than trying to neutralize its effects (i.e., the passage dependency phenomenon). They also included items that were designed to measure students' use of reading strategies and their dispositions toward reading. For example, the Illinois Goal Assessment Program (1991) promoted

TABLE 2.3
A Set of Contrasts Between Cognitively-Oriented Views of Reading and Prevailing Practices in Assessing Reading Circa 1986

New Views of the Reading Process Tell Us That …	*Yet When We Assess Reading Comprehension, We …*
Prior knowledge is an important determinant of reading comprehension.	Mask any relation between prior knowledge and reading comprehension by using lots of short passages on lots of topics.
A complete story or text has structural and topical integrity.	Use short texts that seldom approximate the structural and topical integrity of an authentic text.
Inference is an essential part of the process of comprehending units as small as sentences.	Rely on literal comprehension text items.
The diversity in prior knowledge across individuals as well as the varied causal relations in human experiences invite many possible inferences to fit a text or question.	Use multiple-choice items with only one correct answer, even when many of the responses might, under certain conditions, be plausible.
The ability to vary reading strategies to fit the text and the situation is one hallmark of an expert reader.	Seldom assess how and when students vary the strategies they use during normal reading, studying, or when the going gets tough.
The ability to synthesize information from various parts of the text and different texts is the hallmark of an expert reader.	Rarely go beyond finding the main idea of a paragraph or passage.
The ability to ask good questions of text, as well as to answer them, is the hallmark of an expert reader.	Seldom ask students to create or select questions about a selection they may have just read.
All aspects of a reader's experience, including habits that arise from school and home, influence reading comprehension.	Rarely view information on reading habits and attitudes as important information about performance.
Reading involves the orchestration of many skills that complement one another in a variety of ways.	Use tests that fragment reading into isolated skills and report performance on each.
Skilled readers are fluent; their word identification is sufficiently automatic to allow most cognitive resources to be used for comprehension.	Rarely consider fluency as an index of skilled reading.
Learning from text involves the restructuring, application, and flexible use of knowledge in new situations.	Often ask readers to respond to the text's declarative knowledge rather than to apply it to near and far transfer tasks.

Note. This was adapted from Valencia and Pearson, 1987b, p. 731.

an interactive model of reading in which the construction of meaning became the locus around which reading strategies, dispositions toward literacy, text characteristics, and prior knowledge, revolved. The question in Fig. 2.2 illustrates the possibility of having more than one right answer (coded with a *) to the question.

Another powerful influence focused on the test development process rather than items per se. Consistent with the work on text structure from the early part of the cognitive revolution, many of the new assessments used elaborate text analyses procedures to generate structural representations of the texts (story maps for narratives and text maps for informational texts) to be used in developing comprehension assessments. Equipped with these structural representations, which were also hierarchical, test developers used criteria of structural importance to decide which subset, among the myriad of conceptual relations within a text, ought to be included in a comprehension assessment. Test writers employed this technique as a part of the test specification procedures in several state assessment efforts (Valencia et al., 1989) and in the National Assessment of Educational Progress from 1988 onward (NAEP Reading Consensus Project, 1992).

A Systematic Research Program. A fair amount of research on these new assessment practices was carried out in the 1980s, much of it conducted at the Center for the Study of Reading under the leadership of Valencia and Pearson (Pearson et al., 1990; Valencia & Pearson, 1987a; Valencia et al., 1986). For example, several candidate measures of prior knowledge were compared to a common criterion, an individual interview, to determine which exhibited the greatest concurrent validity (Pearson et al., 1990). This work was a part of a new way of dealing with the prior knowledge problem in reading comprehension assessment. As we mentioned earlier, the traditional approach to dealing with prior knowledge in standardized tests was to neutralize it. Test writers would

What do you think that the author Patricia Edwards Clyne wanted you to learn from reading "The Army of Two"?

A. There is safety in large numbers.
B. Keep things that you may need in the future in a safe place.
C. Lighthouses and sand dunes are dangerous places to live.
*D. It takes more than strength to win a battle.
*E. Careful thinking can sometimes make things possible that seem impossible.

FIG. 2.2. An item that has more than one right answer.

provide lots of short passages covering a wide variety of topics, the hope being that the variety would prevent any given type of individual from being consistently advantaged because of prior experiences.[16] The solution advocated in the 1960s was to use passage dependency analyses as a means of culling out items that could be answered without reading the text. The solution in these new assessments was to embrace prior knowledge as a part of the process of making meaning and then to assess it independently of comprehension so that its impact could be separately indexed.

Similar criterion validity studies were carried out for measures of comprehension monitoring, dispositions for reading, and comprehension. Although this work addressed a broad range of psychometric and conceptual issues, item format and test infrastructure is of greatest interest to the problems still lingering in the field. Central questions still plaguing us are which formats have the greatest validity as indexes of comprehension and how do the various items in a comprehension assessment cluster form independent factors.

The first analysis of interest in the Illinois work is a criterion validity study carried out in the 1986 pilot (Pearson et al., 1990). They investigated the relation between a common interview technique for assessing passage comprehension and four competing multiple-choice assessment formats. The four formats were conventional multiple choice: select as many answers as are plausible, rate each answer on a scale of plausibility, and select that subset of questions that would tap important information in the passage. These formats are described in Table 2.4.

Working with 200 eighth-graders who had taken the test in one of the four formats (50 per format) and operating under the assumption that in a perfect world, a one-on-one interview format would give us the best possible index of any students' comprehension, they conducted a Piagetian-like clinical interview (see Ginsburg, 1997) to interrogate their understanding of the same passage. Using retelling and follow-up question probes, each student received an interview score characterizing the breadth and depth of his or her comprehension. This score was used as the criterion variable to compare the common variance that each format shared with the interview. The researchers hypothesized that if comprehension consists of deep reasoning about passage content, then formats which emphasize deeper processing of content ought to be more closely related to (and hence serve as better predictors of) the ultimate interview criterion than those formats requiring lower levels of processing. The results supported the following hypothesis: the "rate each of the choices" format shared

[16]Note that this approach tends, on average, to favor those students who have high general verbal skills as might be indexed by an intelligence test, for example. These will be the students who will possess at least some knowledge on a wide array of topics (Johnston, 1984a, 1984b).

TABLE 2.4
Description and Examples of Response Formats in the Illinois Pilot of 1986

"Single Correct Answers"

The standard comprehension format was a multiple-choice item where students select the one best answer to a question.

How does Ronnie reveal his interest in Anne?

Ronnie cannot decide whether to join in the conversation.
Ronnie gives Anne his treasure, the green ribbon.
Ronnie invites Anne to play baseball.
Ronnie gives Anne his soda.

"More Than One Right Answer"

The second item format was constructed to look very much like a traditional multiple-choice item but with one important difference: Students are told that there could be one, two, or even as many as three plausible responses to each question. The rationale behind such a format is that most questions do have more than one right answer. The ideas presented in stories and in nonfiction selections often have multiple causes and multiple relations with other ideas in the text or related to the text; hence, it is very constraining, if not misleading, to have to write items that allow only a single explanation for a complex relation.

How does Ronnie reveal his interest in Anne?

Ronnie cannot decide whether to join in the conversation.
Ronnie gives Anne his treasure, the green ribbon.
Ronnie gives Anne his soda.
Ronnie invites Anne to play baseball.
During the game, he catches a glimpse of the green ribbon in her hand.

"Score-Every-Response" Format

A slight variation of this "select-as-many-as-are-appropriate" format was developed for use at Grade 8 and Grade 10. Given items that have one, two, or as many as three plausible answers (as in the previous format), students must score each response option with a 2, 1, or 0 rating, where 2 indicates a *very good answer,* 1 indicates that the response is *on the right track,* and 0 represents a response that is *totally off the track*. In this format, students must deal with every response option, and they must use the defined criteria to help them distinguish levels of appropriateness:

How does Ronnie reveal his interest in Anne?

(2) (1) (0) Ronnie cannot decide whether to join in the conversation.
(2) (1) (0) Ronnie gives Anne his treasure, the green ribbon.
(2) (1) (0) Ronnie gives Anne his soda.
(2) (1) (0) Ronnie invites Anne to play baseball.
(2) (1) (0) During the game, he catches a glimpse of the green ribbon in her hand.

"Question-Selection" Format

In the fourth format, students were presented with a list of 20 questions that might be asked about the passage they read. The task was to select approximately 10 questions that would be good to ask classmates to be sure they understood the reading selection. Students were not to answer these questions, but simply to identify good questions by marking each with a "Yes" or a "No." This item format is based on the research finding that skilled readers are better than poor readers at asking questions, both to clarify confusing points and to focus on important aspects of text.

the most variance in common with the interview score, followed, in order, by the "select-all-the-plausible-answers" format, the "conventional format," and the "select questions" format. Interestingly, the "rate-each-distracter" format also achieved the highest reliability ($\alpha = .93$).

The Illinois group also carried out two important factor analytic studies during the 1987 pilot. 2,700-plus students at each of four grade levels—3, 6, 8, and 10, with each student responding to comprehension, prior knowledge, metacognitive, and habits and attitudes items from two out of six passages, with each pair of passages occurring equally often in a matrix sampling plan. Using exploratory factor analysis, a three-factor solution emerged consistently across all passage pairs. Essentially the metacognitive and the habits and attitudes items each defined an independent factor with the third factor being a combination of the comprehension and prior knowledge items. Given the centrality of prior knowledge in the underlying schema-theoretic models on which the assessment was built, the clustering of knowledge and comprehension items was not a surprise. However, it must be acknowledged that the reliability of the prior knowledge scale, when calibrated at the individual level, was much lower than the reliability of the other tests.

In a second factor analysis, the group investigated whether responses to individual comprehension items across the 16 pairs of passages tended to cluster more by cognitive category or passage. Using a combination of exploratory and confirmatory factor analyses, they were unable to achieve any clustering by cognitive category. For all 16 passage pairs, passage, not cognitive process, emerged as the single explanatory factor. This led them to conclude that topical knowledge, not cognitive process, was a more salient factor in explaining variance in reading comprehension.

Sentence Verification Task. Another novel assessment approach emerging from the cognitive science work of the 1970s was the sentence verification task (SVT). It was developed by Royer and his colleagues (Royer, 1987; Royer, & Cunningham, 1981; Royer, Hastings, & Hook, 1979; Royer, Kulhavy, Lee, & Peterson, 1986; Royer, Lynch, Hambleton, & Bulgarelli,1984) to provide a measure of reading comprehension that was not, like so many multiple-choice standardized tests, dependent on external factors such as intelligence or prior knowledge. They have also championed it as a task that teachers can adapt to assess comprehension of specific texts used in their classrooms. One of its other attributes is that, like the cloze task, SVTs can be created using a procedure that involves relatively few inferences and judgments on the part of the test developer. Once a passage has been selected, item development is quite algorithmic. The test developer selects a passage, such as the one about down pillows.

One wonderful thing about grandparents, Tim decided, was the stories they could tell about his parents when they had been young. His favorite story about his mother was the famous pillow caper.

"Nowadays," Grandma said, "a feather pillow is something of a rarity or a luxury. Most people seem content with polyester fillings and such. When your mother was small, we had nothing but feather stuffed in our house. You don't know what comfort is until you've sunk your head into 3,000 bits of goose down.

"Once when your mother had nothing to do, she saw the point of one little feather sticking out of a tiny hole in the corner of her pillow. She pulled it out and another came right along to take its place. You can image the rest of this story!"

"Yes," laughed Tim, "she pulled out all the feathers."

"I went to her room," said Grandma, "and there I found 3,000 feathers flying around. All your mother could say was: 'I didn't know there would be so many of them!'"

Then one proceeds to develop an approximately equal number of four different item types (Table 2.5).

Examples for all four item types appear in Table 2.6. All, incidentally, are derived from the previous passage. When it is administered, an examinee reads the passage and then responds to the items (selecting old or new for each) without being able to look back at the text. Thus, at least short-term memory is required to complete the task.

Royer and his colleagues have applied the SVT to texts in a number of subject matters and to a diverse array of student populations, including ESL populations. The procedure produces results that meet high psychometric standards of reliability and validity (Royer, 1987). In addition, scores on the SVT are sen-

TABLE 2.5
Item Types and Definitions for New Sentence Verification Task

Item Type	*Definition*
• Original:	Verbatim repetition of a sentence in the passage.
• Paraphrase:	The same meaning as an original but with lots of semantic substitutes for words in the original sentence.
• Meaning change:	Uses some of the words in the passage but in a way that changes the meaning of the original sentence.
• Distracter:	A sentence that differs in both meaning and wording from the original.

TABLE 2.6
A Sample Sentence Verification Task Comprehension Test

Choices	*Items*
Old New	1. Most people seem content with polyester fillings and such. (Original)
Old New	2. You don't know what comfort is until you've sunk your head into 3,000 bits of polyester. (Meaning change)
Old New	3. It is always fun visiting grandparents because they take you someplace exciting, like the zoo or the circus. (Destructer)
Old New	4. Being able to hear stories of when his mom and dad were kids was one of the great things about having grandparents around, Tim concluded. (Paraphrase)
Old New	5. His favorite grandparent was his mother's mother. (Destructer)
Old New	6. In our home, we only had pillows filled with feathers when your mom was a child. (Paraphrase)
Old New	7. "Nowadays," Grandma said, "feather pillows are very common and not considered a luxury." (Meaning change)
Old New	8. His favorite story about his father was the famous pillow caper. (Meaning change)
Old New	9. Once when your mother had nothing to do, she saw the point of one little feather sticking out of a tiny hole in the corner of her pillow. (Original)
Old New	10. "I never guessed there would be this many feathers," was the only thing she could say. (Paraphrase)
Old New	11. You can guess what happened next! (Paraphrase)
Old New	12. "I went out to the yard," said Grandma, "and there I found 3,000 feathers flying around." (Meaning change)
Old New	13. She poked it back in, but another came right along to take its place. (Meaning change)
Old New	14. "Yes," laughed Tim, "she pulled out all the feathers." (Original)
Old New	15. "I wish," said Tim, "that I could get a goose down pillow." (Distracter)

sitive to other factors that are known to affect comprehension, such as prior knowledge (Royer, Lynch, Hambleton, & Bulgarelli, 1984), overall reading skill (Royer & Hambleton, 1983), intersentential comprehension (Royer, Kulhavy, Lee, & Peterson, 1984), and text readability (Royer et al., 1979). Despite a good track record and strong grounding in both the psychometric and conceptual poles, SVT never gathered much momentum in the field. We suspect that for many educators, it flunks the prima facie test: It just does not have the look and

feel of what we mean by "comprehension assessment." After all, there is no retelling and no question answering. This lack of interest is unfortunate because the technique, or at least some of its features, could be useful in building new, conceptually sound, efficient, and replicable assessment procedures.

Classroom Assessment. The most significant advances in classroom comprehension assessment tools during this period also came from cognitive science. First was the spread of retellings as a tool for assessing comprehension. Driven by the 1970s advances in our knowledge about the structure of narrative and expository text (see Meyer & Rice, 1984), many scholars (see Irwin & Mitchell, 1983; Morrow, 1988) developed systems for evaluating the depth and breadth of students' text understandings based on their attempts to retell or recall what they had read. Like the formal efforts of this era, there was a conscious attempt to take into account reader, text, and context factors in characterizing students' retellings.

Second was the "use the think-aloud" protocol as a measure of comprehension. Think-alouds had become respectable research tools by virtue of the important work on self-reports of cognitive processes popularized by Ericsson and Simon (1984). In attempting to characterize the nature of expertise in complex activities, such as chess, Ericsson and Simon learned that the most effective way inside the heads of expertise was to engage the players in thinking aloud about the what, why, and how of their thing and actions during the activity.

This led to the wider use of think-alouds. First, they became a research tool to get at the process, not just the product of student thinking (e.g., Hartman, 1995; Olshavsky, 1976–1977). Then, they became an instructional practice (Baumann, Jones, & Seifert-Kessell, 1993), and finally, it was used as an assessment practice (California Learning Assessment System, 1994; Farr & Greene, 1992). With the ostensible purpose of assessing metacognitive processes during reading, Farr and Greene (1992) engaged students in write-along tasks (a kind of mandatory set of marginal notes prompted by a red dot at key points in the text). Students were encouraged, as they are in think-alouds, to say (in this case, make a few notes about) what they thought at a given point. A similar practice was a standard part of the now defunct California Learning Assessment System (1994): marginal notes were allowed, even encouraged, in the initial reading of the texts, and those notes were fair game for review when the tasks were scored. Unfortunately, with the exception of a very thorough account of the research and theoretical background on verbal protocols by Pressley and Afflerbach (1995), very little careful work of either a conceptual or psychometric nature on the use of think-alouds as a viable assessment tool has emerged, although there was one effort to evaluate different approaches to metacognitive assessment in

the special studies of the National Assessment of Educational Progress (NAEP) in 1994; in fact, this effort spawned the Farr and Greene effort.

SOCIOCULTURAL AND LITERARY PERSPECTIVES

We are not sure whether what happened next constitutes a second major shift or is better thought of as an extension of the first shift. It came so fast on the heels of the cognitive revolution that it is hard to pinpoint its precise beginning point.

Sociolinguistics

In fact, harbingers of this sociocultural revolution, emanating from sociolinguistic perspectives (see Bloome & Greene, 1984) and the rediscovery of Vygotsky (see Vygotsky, 1978; Wertsch, 1985), were around in the early to mid-1980s, even as the cognitive revolution was exercising its muscle on assessment practices. For example, in cognitively motivated teaching approaches such as reciprocal teaching, students took on more responsibility for their own learning by teaching each other. In process writing, revision, and conversation around revision, delved more deeply into the social nature of reading, writing, and understanding. Teachers used such practices to engage students to reflect on their work as well as interact with others around it. The concept of "dynamic assessment" also emerged in this period. Dynamic assessment (Feuerstein et al., 1979) allows the teacher to use student responses to a given task as a basis for determining what sort of task, accompanied by what level of support and scaffolding from the teacher, should come next. Here we see both cognitive and sociocultural influences in assessment.

These early developments notwithstanding, the next round of assessment reforms carried more direct signs of the influence of these new social perspectives of learning, including group activities for the construction of meaning and peer response for activities requiring writing in response to reading.

Literary Theory

The other influential trend was a renaissance in literary theory in the elementary classroom. One cannot understand the changes in pedagogy and assessment that occurred in the late 1980s and early 1990s without understanding the impact of literary theory, particularly reader response theory. In our secondary schools, the various traditions of literary criticism have always had a voice in the curriculum, especially in guiding discussions of classic literary works. Until the middle 1980s, the "New Criticism" (Richards, 1929) that began its ascen-

dancy in the depression era dominated the interpretation of text for several decades. It had sent teachers and students on a search for the one "true" meaning in each text they encountered.[17] With the emergence (some would argue the reemergence) of reader response theories, all of which gave as much authority to the reader as to either the text or the author, theoretical perspectives, along with classroom practices, changed dramatically. The basals that had been so skill-oriented in the 1970s and so comprehension-oriented in the 1980s became decidedly literature-based in the late 1980s and early 1990s. Comprehension gave way to readers' response to literature. Reader response emphasizes affect and feeling that can either augment or replace cognitive responses to the content. To use the terminology of the most influential figure in the period, Louise Rosenblatt (1978), the field moved from efferent to aesthetic response to literature. And a "transactive model" replaced the "interactive model" of reading championed by the cognitive views of the 1980s. According to Rosenblatt, meaning is created in the transaction between reader and text. This meaning, which she referred to as the "poem," is a new entity that resides above the reader-text interaction. Meaning is therefore neither subject nor object nor the interaction of the two. Instead, it is transaction, something new and different from any of its inputs and influences.[18]

Illustrating the Impact of Reading Assessment

Nowhere was the influence of these two new perspectives more prominent than in the development of the California Language Arts Framework (California Department of Education, 1987) and in the assessment systems that grew out of the framework. There was a direct attempt to infuse social, cultural, and literary perspectives into comprehension assessment processes more transparent than in the work of the California Learning Assessment System (CLAS; 1994). CLAS, which died an unhappy death via legislative mandate in the mid-1990s, nonetheless paved the way for more open assessments by emphasizing response to literature formats and the social aspects of learning. Response to literature questions articulated a more open and reflective stance toward reading rather than a skills-based approach:

- If you were explaining what this essay is about to a person who had not read it, what would you say?

[17]We find it most interesting that the ultimate psychometrician, Frederick Davis (e.g., 1968), was fond of referencing the New Criticism of I. A. Richards (1929) in his essays and investigations about comprehension.

[18]Rosenblatt (1978) credited the idea of transaction to John Dewey, who discussed it in many texts, including *Experience and Education* (1938).

- What do you think is important or significant about it?
- What questions do you have about it?
- This is your chance to write any other observations, questions, appreciations, and criticisms of the story" (pp. 6–9).

Response to literature formats demanded students to be able to summarize, explain, justify, interpret, and provide evidence in their answers. In other words, assessment of reading comprehension reached a new stage, one much more compatible with what society might expect of students in the real world. The early work of the New Standards (see Pearson, Spalding, & Myers, 1998) had the same goals, theoretical grounding, and format characteristics as CLAS (1994):

- Give students a chance to show their expertise in artifacts that have the benefit of the same social and cultural supports that support effective instruction.
- Let the work be interesting and relevant to students? backgrounds and cultural heritage.
- Let the work be guided by the support of colleagues who have the students' best interests at heart.
- Let the work be borne of the same motives and conditions that prevail in the worlds of work and social action.

If the idea that students live in multiple worlds such as home, school, and community and are expected to relate to others across contexts grew out of the sociocultural revolution in the late 1980s, it must be acknowledged that this revolution had well-grounded historical precedents (see Dewey, 1938). In line with the idea of the social nature of learning, comprehension assessment systems such as CLAS and New Standards also devised reading comprehension tests which focused on the interconnectedness of individual learning within the contexts of group work. Figure 2.3 is such an example.

These changes in reading assessment practices did not go unnoticed by basal reader publishers. Beginning in the late 1980s and early 1990s, they began to incorporate these practices into their internal assessment systems, using handles such as process tests and performance assessments (see, for example, Silver Burdett & Ginn, 1989). Basals maintained their specific skill tests, but even with these carryovers from the 1970s, important changes occurred—longer passages, assessment of multiple skills per passage, and a reduction in the number of skills assessed. By the mid-1990s, basal assessments had moved even further down the performance assessment road. Even so, they never completely eliminated the specific skill tests; instead, these new assessments were added as optional alternatives to the more traditional tools.

Now you will be working in a group. You will be preparing yourself to do some writing later. Your group will be talking about the story you read earlier. A copy of the story is provided before the group questions if you need to refer to it. Some of the activities in this section may direct you to work alone and then share with your group, the other activities may have all of you working together. It is important to take notes of your discussion because you will be able to use these notes when you do your writing.

Read the directions and do the activities described .members of the group should take turns reading the directions. The group leader should keep the activities moving along so that you finish all the activities.

You'll have 15 minutes for these prewriting activities.

FIG. 2.3. An activity from the California Learning Assessment System (1984).

Critiques of the New Assessments

As with other novel approaches in comprehension assessment, performance assessments came under fire as teachers and test developers struggled with issues of validity (particularly for individual scores), external accountability, reliability, and generalizability (Linn, 1999; Pearson, DeStefano, & García, 1998). Given what we know about the high stakes functions for which assessments are used to make decisions about individuals (e.g., decisions about entry into or exit from special programs or "certifying" or licensure decisions), these criticisms should not be surprising.

The Social Nature of the Assessments. Performance assessments, probably because of their strong connection to everyday classroom activity and real-world workplace contexts, tended to encourage teachers to have students work in groups. This led to an essential dilemma: What are we to do when we know that the performance of an individual student is influenced by the work, comments, and assistance of peers or teachers? The essence of this dilemma was captured well in an essay by Gearhart et al. (1993) entitled, "Whose work is it anyway?" This "contamination" of individual student scores has prompted great concern on the part of professionals who need to make decisions about individuals. The social components of the reading process can be grounded in theories that may even deny the existence, or at least the significance, of the "individual." This makes assessment doubly difficult.

Task Generalizability. Task generalizability, the degree to which performance on one task predicts performance on a second, is a major concern with these performance tasks. The data gathered from the first scoring of New Standards tasks (Linn, DeStefano, Burton, & Hanson, 1995) indicate that indexes of generalizability for both math and reading tasks were quite low. That essentially means that performance on any one task is not a good predictor of scores on other tasks. Shavelson and his colleagues encountered the same lack of generalizability with science tasks (Shavelson, Baxter, & Pine, 1992), as have other scholars (e.g., Linn, 1993), even on highly respected enterprises such as the advanced placement tests sponsored by the College Board. The findings in the College Board analysis are noteworthy for the incredible variability in generalizability found as a function of subject matter. For example, to achieve a generalizability coefficient of .90, estimates of testing time range from a low of 1.25 hr for Physics to over 13 hr for European History. These findings suggest that we need to measure students' performance on a large number of tasks before we can feel confident in having a stable estimate of their accomplishment in a complex area such as reading, writing, or subject matter knowledge. Findings such as these probably explain why standardized test developers have included many short passages on a wide array of topics in their comprehension assessments. They also point to a bleak future for performance assessment in reading; one wonders whether we can afford the time to administer and score the number of tasks required to achieve a stable estimate of individuals' achievement.

The Legacy. If one examines trends in the assessment marketplace and in state initiatives, one can make predictions based on a usually reliable indicator of the latest trends in assessment. Now, in the year 2004, the revolution begun in the 1980s is over, or at least inching along in a very quiet cycle. Granted, successful implementations of authentic wide-scale assessment have been maintained in states like Maryland (Kapinus, Collier, & Kruglanski, 1994), Kentucky, and Oregon (see Pearson, Calfee, Walker-Webb, & Fleisher, 2002). However, other states (e.g., California, Wisconsin, Arizona, and Indiana) have rejected performance assessment and returned to off-the-shelf, multiple-choice, standardized reading assessments. Pearson et al. (2002) found a definite trend among states in which performance assessment is still alive to include it in a mixed model, not unlike NAEP, in which substantive, extended response items sit alongside more conventional multiple-choice items. Both these item formats accompany relatively lengthy passages. Even the more modest reforms in Illinois (the multiple-correct answer approach) were dropped in 1998 (interestingly, in favor of a NAEP-like mixed model approach). And it is the NAEP

model that, in our view, is most likely to prevail. It is within the NAEP mixed model that the legacy of the reforms of the early 1990s are likely to survive, albeit in a highly protracted form. It is to the NAEP experience that we now turn our attention.

THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

It is important to give a prominent role to the reading assessments of NAEP in examining the history of reading comprehension assessment in the last quarter of the 20th century. NAEP is often regarded by people in the field as a sort of gold standard for the most forward thinking assessment practices (see Campell, Voelkl, & Donahue, 1998). NAEP has, especially in the last decade, developed a reputation for embodying our best thinking about both the conceptual and psychometric bases of assessment, especially in the area of reading.

The History of NAEP Reading Assessment

NAEP was authorized by Congress in 1963 as an effort to assess the "condition and progress of American education." The first NAEP assessment was administered in 1969, and it has grown in stature and importance from one decade to the next. Overseen by the National Assessment Governing Board and the National Center for Educational Statistics (Jones, 1996), the NAEP tests are the grades on the infamous "Nation's Report Card," which serves as an index of how states compare. It is also an index of trends in achievement.

The original plan for developing the NAEP tests was to center on open-ended constructed response items, which were to be interpreted within a "goal-free" perspective; that is, there would be little attempt to aggregate or interpret scores across items to build profiles of subdomains within a subject matter. Instead, performance data would be tied to individual items, which would be made public; the hope was that the significance of various performance data would emerge from these public forums. Two things changed quickly, however. First, for whatever reasons (economy being the most plausible), multiple-choice, not constructed response, dominated even the early years of NAEP, and they still constitute the lion's share of items (with allotted time balanced fairly equally between multiple-choice and constructed response). Second, in the early 1970s, NAEP moved quickly from its goal-free position to adopt subject matter frameworks, with the clear expectation that items would be developed to provide an overall account of performance in subject matter domains and to measure specific objectives within the framework.

In reading, revisions of frameworks have prompted changes in the reading test instrument over time, at least since the early 1970s (Salinger & Campbell, 1998). In the framework operative in the 1970s, student performance was measured against a set of objectives that looked remarkably consistent with then popular Bloom's taxonomy (1956). Students should

- Demonstrate the ability to show comprehension of what was read
- Analyze what is read, use what is read.
- Reason logically.
- Make judgments.
- Have an attitude and interest in reading.

The 1980s framework reveals several changes, some of which indicate that literature was beginning to make its mark on the objectives. Students would, according to the framework

- Value reading and literature.
- Comprehend written works.
- Respond to written works in interpretive and evaluative ways.
- Apply study skills.

In 1992, amidst the renaissance in literature as a part of the elementary and secondary curriculum, the impact of reader response theory on the NAEP framework is transparent. In fact, the "stances" readers are supposed to take are directly traceable to Langer's approach to helping teachers implement a response-based curriculum (see Langer, 1995). Langer's approach to response includes activities that get students into, through, and beyond the texts they read. In the 1992 NAEP framework, students were to be exposed to texts and items that encourage them to read to

- Form an initial understanding.
- Develop an interpretation.
- Personally reflect and respond to the reading.
- Demonstrate a critical stance.

Forming an initial understanding is remarkably like Langer's (1995) "into" stage. Developing an understanding is what Langer had in mind in her "through" stage, and personal response and critical reflection map directly onto Langer's "beyond" stage.

The Current NAEP Framework

The stances are the driving force behind the latest framework. It is worth dissecting these categories to examine what can be thought of as a theoretically well-grounded approach to item development. "Forming initial understanding" focuses on readers' initial impressions or "global understanding" (p. 11) of the text, with a broad perspective. NAEP often puts this as one of the first questions on a test. "Developing interpretation" occurs when readers are required to "extend their initial impressions to develop a more complete understanding of what they read" (p. 11) by taking information across parts of a text and focusing on specific information. In "personal response," readers are required to "connect knowledge from the text with their own personal background knowledge." In other words, how does the text relate to personal knowledge? In the last stance, "demonstrating critical stance," readers are required "to stand apart from the text and consider it objectively" (p. 12). This involves "critical evaluation, comparing and contrasting, and understanding the impact of such features as irony, humor, and organization" (p. 12). These stances are illustrated by several sample questions taken directly from the 1992 Framework booklet (see Table 2.7; NAEP Reading Consensus Project, 1992).

Neither the NAEP Reading Framework nor the NAEP item structure has changed since 1992.[19] The framework and the items are designed to allow us to understand how well American students negotiate complex interactions with text. This goal is to be achieved by the inclusion of a wide variety of text types and purposes (reading for literary experience, reading for information, and reading to perform a task), strict attention to the "four stances" described earlier, and systematic inclusion of three item types: multiple choice, short constructed response, and extended constructed response.

Issues and Problems in NAEP Reading Assessment

The Framework. Despite the efforts of those responsible for NAEP to ensure that the reading framework reflects the latest knowledge gleaned from recent research, many criticisms abound. For example, the four stances can overlap, both conceptually and psychometrically. A simple examination of the questions suggests that often personal response and developing interpretation are a part of critical reflection; in getting to a critical position, a reader often ex-

[19] As this manuscript was completed, no change in the National Assessment of Educational Progress (NAEP) Reading Framework (NAEP Reading Consensus Project, 1992) had surfaced, although a task force had been assembled to consider changes.

TABLE 2.7
National Assessment of Educational Progress (NAEP) Stances

Forming Initial Understanding

- Which of the following is the best statement of the theme of the story?
- Write a paragraph telling what this article generally tells you.
- What would you tell someone about the main character?

Developing Interpretations

- How did the plot begin to develop?
- What caused the character to do this (use examples from the story to support your answer)?
- What caused this event?
- What type of person is the character (use information from the text to support your answer)?
- In what ways are these ideas important to the topic or theme?
- What will be the result of this step in the directions?
- What does the character think about ____?

Personal Reaction and Response

- How did this character change your ideas of _____?
- Do you think that ___ (say a grandmother or a child) would interpret this passage in the same way?
- How is the story like or different from your own personal experience? Explain?
- Does this description fit what you know about ___ and why?
- What does this passage/story say to you?
- Why do you think ___ (bullfrogs eat dragonflies? Is there anything else you think they might eat? What information from your own knowledge helps you answer this)?

Demonstrate Critical Stance

- Compare this article/story to that one.
- How useful would this be to ____?
- Do you agree with the author's opinion of this event?
- Does the author use (irony, personification, humor) effectively? Explain.
- What could be added to improve the author?s argument?
- Is this information needed?
- What other information would you need to find out what you don't know now?

amines both his or her knowledge repertoire and reads across text segments. It is interesting to note that NAEP has never reported scores by these stances, although the sample of items per category in their matrix-sampling scheme is large enough to obtain reliable estimates for each of the categories. The most plausible conclusion is that the cognitive categories do not hold up when scrutinized by factor analysis and other techniques designed to empirically determine

the patterns of internal clustering. Furthermore, even expert literacy researchers cannot reliably classify items into the categories assigned by NAEP test developers (DeStefano, Pearson, & Afflerbach, 1997). Essentially, researchers judged the four types not to be discrete from one another. This failure of the cognitive stances to hold up psychometrically is reminiscent of the findings from the analyses of the Illinois assessment a decade earlier.

Item Format. Perhaps the most serious validity issue regarding test items within NAEP centers on item format. In particular, the push for achieving greater economy of testing time and scoring costs have prompted NAEP officials to fund research on the "value added question" of constructed response items; the question of interest is whether the extra expense of constructed response items is justified. That expense would be justified if and only if it could be demonstrated that constructed response items increase our capacity to assess comprehension accurately and validly—above and beyond what could be achieved only with multiple-choice items. Evidence of increased capacity could take many forms: (a) the discovery of independent factors for different formats, (b) different item information functions (a la item response theory), and (c) finding that different items elicit different levels of cognitive engagement.

Focusing on the third of these potential sources of evidence, Pearson et al. (in press) conducted several studies to determine if there are any substantive differences in cognitive processes evoked from multiple-choice and constructed-response items with the 1996 NAEP reading data. The basic methodology involves asking students to think aloud as they answer questions. Using both the item responses and the think-aloud data, researchers classify the type of thinking students were engaging in as they worked through each answer. The think-aloud produces cognitive categories, which are then scaled on a "cognitive depth" continuum. These depth indexes are used to compare depth across different item formats.

In the initial study, which was limited to existing NAEP passages and items, the research team members found few cognitive differences across the multiple-choice and constructed-response divide. Concerned that they might not have captured depth of processing adequately, they went outside the NAEP framework and item sets. First, they reframed the concept of deeper engagement using a more theoretically driven lens. Then they created the best possible test for "open-ended" constructed response items in reading by including a text type not previously studied, poetry, and tasks that were very much in the reader response tradition, discussed earlier. In addition, they wanted to see if the presence of multiple texts could also influence engagement, so they chose three thematically related poems and adapted some very open-ended constructed-response items that

were much more in the spirit of the CLAS assessments discussed earlier. In scoring the new think-alouds, they discovered two new indexes of deep engagement that had not emerged in the earlier analysis (which was limited to NAEP tasks)—"multiplicity" and "intertextuality." Multiplicity occurs when an examinee assumes more than one possible stance toward an item; for example, taking the perspective of the author, then the main character, then perhaps himself or herself as a reader. Intertextuality involves linking ideas across texts (or distinctly separate segments of a single text). With these new tools for examining deeper cognitive engagement, they re-analyzed the earlier NAEP think-alouds only to discover that the data from those tasks also exhibited these two new indexes of depth, prompting them to re-analyze the earlier data set. The re-analysis indicated that the multiple-choice items elicited a significantly lower proportion of multiple and intertextual strategies than did either the short or extended constructed-response items. These data suggest that when the standard of deep engagement is relatively low, as it was in the first study, few item format differences emerge, but that when the bar for deep cognitive engagement is set higher (number of think-aloud statements exhibiting either multiplicity of stances or textual linking), item format differences emerge. Moreover, the data from the poetry task suggest that if we really want to assess deeper comprehension processes, we are better advised to develop genuine performance tasks than we are to simply transform multiple-choice items into short answer questions.

OTHER IMPORTANT DEVELOPMENTS IN THE LAST DECADE

Linking Comprehension Assessment to Book Reading Levels

Within the last decade, two separate initiatives have been tried to link performance on tests of reading comprehension to books that students ought to be able to read. Both the Degrees of Reading Power (Touchstone Applied Science Associates, 1995) and the Lexile scales (Stenner & Burdick, 1997; Stenner et al., 1987) provide this sort of service. They are able to achieve this link by placing students' scores on comprehension measures on the same scale as the readability of books and other reading materials. Scores on a particular test, then, indicate more than how an examinee's performance compares to other examinees (norm-referenced) or to some preset cutoff score (criterion-referenced). Additionally, they point to the sort of books that a student achieving a particular score ought to be able to read. Specifically, they indicate the level of difficulty (what some might call challenge) of books that students achieving a given score ought to be able to read and understand (i.e., answer 75% of a hypothetical set of com-

prehension questions that might be asked about the book). One might think of Lexile scores as "text referenced" measures of comprehension, although Stenner and Burdick (1997) refer to this feature as criterion-referenced.

To validate the Lexile framework, Stenner and his colleagues (Stenner & Burdick, 1997; Stenner et al., 1987) have engaged in an elaborate sequence of studies. Using the mainstays of readability formulas (word frequency and sentence length) as a way of predicting the difficulty of text, they scaled the difficulty of a wide range of cloze-like comprehension test passages and items, as illustrated in Fig. 2.4.

After administering the test items to a large set of examinees and validating the predictive power of their indexes of passage difficulty, Stenner and his colleagues created a set of formulas that allowed them to place examinee performance and text difficulty on a common scale: "An important feature of the Lexile Framework is that it provides criterion-referenced interpretations of every measure. When a person's measure is equal to the task's calibration, then the Lexile scale forecasts that the individual has a 75 percent comprehension rate on that task. When 20 such tasks are given to this person, one expects three-fourths of the responses to be correct" (Stenner & Burdick, 1997, p. 16).

Stenner and his colleagues then applied the Lexile scaling techniques to 1,780 comprehension items from nine popular standardized tests, obtaining average correlations between Lexile predictions and observed item difficulties in the mid-.90 range. The next major initiative was to validate the Lexile as a measure of text difficulty, which they accomplished by using the scale to predict the rank ordering of basal reading levels (preprimer through Grade 8 reader); the average correlation, across 11 basal series, turned out to be .97.

The final step is to apply the lexile scale to a wide range of trade books, which the Stenner group has done. Now it is possible for a teacher to receive scores from a standardized test in Lexiles as well as National Curve Equivalents (NCEs), or

An Example Lextile Test Item

Wilbur likes Charlotte better and better each day. Her campaign against insects seemed sensible and useful. Hardly anybody around the farm had a good word to say for a fly. Flies spent their time pestering others. The cows hated them. The horses hated them. The sheep loathed them. Mr. and Mrs. Zuckerman were always complaining about them, and putting up screens. **Everyone _____ about them.**

A. agreed B. gathered C. laughed D. learned

FIG. 2.4. Segment from *Charlotte's Web* by E. B. White, 1952, New York: Harper & Row.

percentiles, or grade-norms (for example, both the Stanford Achievement Test–9 and the Stanford Diagnostic Reading Test from HBJ provide Lexile score reporting options). The idea is that teachers can then use the scores to guide them in helping students select appropriate books (at least for the subset of books that have been "Lexilized").

The Degrees of Reading Power (DRP), a modified cloze test originally developed by the state of New York for state assessment purposes, has developed a similar approach to scaling books and students on a common scale, allowing teachers to use DRP scores to place students in level appropriate books. A major difference between the DRP approach and the Lexile approach is that although the DRP scale requires the use of its test, the Lexile scale can be (and has been) applied to any of several currently available standardized tests. What we have been unable to locate is any research indicating the accuracy and validity of the book matching process (e.g., some independent measure of whether the books predicted as within a child's range really were), save a few Web site testimonials from customers.

Reemergence of Skills Orientation

After a decade in which reading skill instruction was backgrounded in deference to literature-based activities that took center stage in reading instructional practices, skills have made a remarkable recovery in the past 3 years. All of the state frameworks emerging in the last few years give skills, particularly phonemic awareness and phonics skills, a prominent role, especially in the primary grades. Also, basal programs that only 7 years ago tried to hide their skills now place them prominently in student texts, workbooks, and teacher manuals. What remains to be seen is how this shift toward greater and more prominent skill inclusion will impact comprehension assessment. Will it usher in a renaissance in skills management systems and lots of specific component skill tests, such as those that characterized the criterion-referenced assessments of the 1970s? Or will the assessments in the next decade continue to carry traces of the performance assessment movement of the 1990s?

NEW INITIATIVES

Having traversed this complex and multifaceted landscape of reading comprehension assessment, we conclude this essay by offering a set of recommendations for future initiatives in this important curricular topic. These recommendations are based on our reading of the full history of reading comprehension assessment in the 20th century. Sometimes the recommendations

are derived from a perceived deficit in the research (we just have not examined the question with sufficient care). Some recommendations are based on our professional judgment that it is time to revisit a question or an issue that, although carefully investigated in earlier periods, deserves a fresh look.

Interactions of Ability and Other Factors

No question is more important to address than the question of whether assessments are equally sensitive to student performance at all levels of the achievement continuum. It is possible, for example, that one approach to comprehension assessment provides a better index of comprehension for students who are still struggling with decoding and word recognition whereas another is more appropriate for students who have word level skills under automatic processing control. To take a simple example, we know that for many younger readers, as well as struggling writers and spellers at any age, the requirement to compose written responses will interfere with our capacity to obtain valid estimates of their comprehension. A number of initiatives seem appropriate at this point in time.

Ability and Response Medium. For years, we have asked students to write in response to reading, generally regarding it as a useful approach to comprehension assessment for more complex reading tasks, such as critical response to literature and critical evaluation of ideas and arguments. Yet we also know that the writing requirement can obscure some children's ability to understand and interpret text because of their poor motor skill development, inadequate spelling, and underdeveloped writing dispositions. Also, it is not unreasonable to hypothesize that certain response media differentially affect students of different abilities or proclivities. Some students, especially those who achieve well, might better show critical dispositions on paper; others, for whom writing is a chore, might shine in an oral response mode. We need studies to evaluate this potential interaction between ability and the medium of response. We could build test forms in which response mode and task complexity are systematically crossed, and then administer the different forms to populations with known reading and writing capacities.

Ensuring That Tests Measure the Entire Range of Performance. In 1992, when NAEP began to use achievement levels (Below Basic, Basic, Proficient, and Advanced) to report student achievement, a problematic feature of the NAEP assessment was unearthed. NAEP, like most state assessments, is given only at a few grades (4, 8, and 12), and when passages are selected for inclusion at a given grade level, test developers try to select passages that are

"appropriate" for the grade level. Granted, test developers seek some variability in passage difficulty; however, the range of difficulty on passages remains fairly narrow. When it was announced that 40% of fourth graders in the United States scored "below basic," there were two possible explanations. Either our fourth graders truly perform that poorly, or the test is insensitive to performance differences at the low end of the achievement scale (in other words, the test has no "floor"). The issue of passage difficulty in reading, particularly its potentially depressing effect on performance of students at the lower end of the performance continuum, has been emphasized in a number of recent reports (e.g., Glaser et al., 1997). It has prompted scholars and policymakers to call for the production of easier blocks of NAEP reading items so that low-performing students can at least "make it onto the scale," or in the language of information value, so that we possess more information about the performance of low-performing students.

NAEP (as well as state assessment efforts relying on grade level passages) should be encouraged to include several blocks containing some very easy (perhaps appropriate for one or two grades below the target grade) passages to see if the distribution in the lower end of the scale can be spread out and measured more reliably and sensitively. In that way, greater meaning could be attached to terms like *below basic* (on NAEP). Interestingly, standardized testing companies have long recognized this problem and accommodated it by providing a range of passage difficulty for each form and level of the test. We suspect, however, that many state tests, because of the emphasis on grade appropriate passages, suffer the same problem encountered in NAEP. Ironically, with the Item Response Theory (IRT) models used in today's wide-scale assessment, giving lower ability students easier passages does not provide them with an unfair advantage. It just gives them a greater opportunity to land somewhere on the underlying performance scale of the test. These issues are especially relevant in dealing with aggregated scores; the reporting of individual scores is another matter.

Achievement and Skill Profiles. Just after midcentury, a few studies were conducted addressing the question of what it means to achieve at a particular level on a standardized test, say a particular grade norm score, percentile rank, or stanine. The methodology was remarkably simple but ingenious. Students took one or another standardized tests; at the same time they took a battery of specific skill tests (phonics, vocabulary, and comprehension subskills). Performances on the various skill tests were examined to build different skill profiles. The question of interest was whether those students who score within a given band of performance on a standardized reading comprehension test would exhibit similar skill profiles. In the few studies we were able to find (e.g., Naylor,

1972), the statistical tendency has shown great profile variability within a given band of performance on a comprehension test. What this suggests is that readers are using some sort of compensatory mechanisms. In other words, some readers may achieve a given level of comprehension performance by relying primarily on a rich reservoir of vocabulary knowledge to compensate for underdeveloped decoding and word recognition skills or comprehension strategies. Others may rely primarily on excellent decoding skills, test-wiseness, or comprehension strategies to compensate for weaknesses elsewhere. The compensatory hypothesis contrasts with a common alternative—the notion that some minimal threshold must be achieved on each and every prerequisite skill for comprehension to take place. Given the recent revival in skills-oriented reading instruction, studies of this ilk would be most timely. We would also be able to better address the question of the multifaceted nature of comprehension and, equally important, the relation among decoding, vocabulary, and comprehension skills. Recently, Riddle-Buly and Valencia (2002) conducted a study along these lines. First, they identified a population of "low achievers" from the Washington statewide assessment (i.e., those who scored "below the bar"—levels 1 and 2 out of 4—on-grade level reading performance). Then they administered a battery of language, vocabulary, and reading subskill assessments (e.g., phonics, word identification, and the like) to these students. What they discovered is that there are indeed many, many ways to fall below the bar. In fact, they identified several "profiles" or clusters of readers who differed dramatically in their skill infrastructures—word callers (students who decode words accurately, even automatically, but don't comprehend well), slow word callers (like their word caller counterparts but are not automatic in recognizing words), word stumblers (accurate but slow and dysfluent readers), and slow and steady comprehenders (students who eventually get the meaning but read slowly and deliberately).

Accommodations for Special Populations. Here is an essential question: How much of an accommodation for special populations can be made before crossing the validity line and invalidating the construct the test is designed to measure? The most radical accommodation on a reading test is, of course, reading the passage and the items to the student. On the face of it, this accommodation changes a reading comprehension test into a listening comprehension assessment. The least radical accommodation, more time, probably does little to invalidate the test or the construct because most comprehension tests are designed to be untimed. Between these two extremes lies a range of popular accommodations that may, to one degree or another, erode or compromise the validity of a test. For example, what is the impact of providing a glossary or dic-

tionary? What about reading the test items but requiring students to read the text on their own? Given the increased emphasis on full participation of all students in group assessments, it seems important to address a full range of possible accommodations, making sure to weigh increased participation against potential sources of invalidity.

Item Format

Although there exists a small corpus of careful studies that allow us to examine the relation between multiple-choice and constructed-response items, we still have a great deal to learn. Much of the problem in interpreting the current set of studies is that the research has been more opportunistic than intentional. In the prototypic study, researchers take advantage of an existing test or battery that happens to include both constructed-response and multiple-choice formats. Much rarer are studies in which the researchers have set out to evaluate both the underlying constructs and the validity of the test(s) designed to measure those constructs.

What is needed is a newer examination of the relations between multiple-choice and constructed-response items. Short of a complete evaluation of the item format construct, there are a number of useful initiatives that would allow us to answer the question of value added for performance items with greater assurance than is currently possible.

The Cognitive Demands of Multiple-Choice and Constructed-Response Items. The work of Pearson et al. (in press) and Campell (1999) has provided us with some estimate of the differential cognitive processes that examinees employ in response to different item formats, and we support more work using the basic methodology of the think-aloud verbal protocols. Although think-aloud methodology appears promising for this sort of initiative (Campbell, 1999; Pearson et al., in press; Yepes-Bayara, 1996), it is by no means the only index of cognitive functioning that we should consider. When tasks involve text reading and response, both eye-movement methodology and computer controlled text search (look-back) methodology could provide rich information about the influence of item format on the role of text in selecting and constructing responses as a part of comprehension assessment.

Rubric Research. We have placed rubric research in the format category because rubrics are unique to a particular item format—constructed-response items. Rubrics such as those used in NAEP for scoring constructed-response

items in reading have been roundly criticized. It would be useful to work with NAEP or NAEP-like reading passages to evaluate different approaches to rubric development. The current NAEP rubrics are viewed by critics as too quantitative and only marginally related to the NAEP framework for reading (DeStefano et al., 1997). High dividends might result from a modest investment in creating new rubrics that are driven by the framework and then comparing the quality of information, both psychometrically and pragmatically, received when items are scored by these rubrics versus conventional rubrics. In another vein, we might examine the conceptual genesis of rubrics, paralleling Fredericksen's (1984) questions about whether transforming multiple-choice items into performance items is the same as transforming performance into multiple-choice items. Suppose the rubric for a set of constructed-response items is based on the same conception of underlying dimensions (the psychological construct) as were used to guide the development of a comparable set of multiple-choice items. Such a practice might, in fact, be a reasonable control if our goal is to examine trait equivalence across item formats; however, this practice can also constrain our thinking about the range of possible traits that might be assessed with the constructed-response format and teased out by an appropriate rubric. In other words, in achieving control for conceptual equivalence, we might be losing our capacity to uncover a large set of possible dimensions of the construct that can only be tapped by the constructed-response format. This issue could be addressed in a study in which competing rubrics were developed and used to score a common set of constructed-response items. The first rubric would be developed using a framework initially used to generate multiple-choice items and then extended to constructed-response items. The second rubric would result from a fresh perspective: subject matter experts would be asked to generate a framework and related rubrics for an assessment system consisting only of constructed-response items. The question of interest is whether the two rubrics would yield equivalent scores and or trait information.

Prior Experience and Test Format. The reader's prior experience has two possible realizations, one at the classroom-school level and one at the individual level. At the classroom-school level, it is instantiated as instructional experience (opportunity to learn). If we can be sure of the type of curricular experiences different students have experienced (e.g., emphasis on critical stance or response to literature in reading or basic decoding skills), and if we can locate populations with different curricular histories, we can test constructed-response items under optimal conditions. We can ask the following: Do students who have learned what the items are designed to measure perform at higher levels than students who have received other curricular em-

phases? It would be interesting, for example, to conduct a think-aloud study (along the lines of the work by Pearson et al., in press) in sites that exhibit just such a curricular contrast. At the individual level, of course, prior experience is instantiated as prior knowledge, and its impact is well documented in reading and writing assessment. It would be useful to know whether students provide more elaborate and more sophisticated responses to constructed-response prompts when they are quite knowledgeable about the topic at hand.

Transforming Items Across Formats. When evaluating the equivalence of constructed-response and multiple-choice items, researchers sometimes begin with one set of items, say multiple-choice, and rewrite them as constructed-response, or vice-versa. In this way, they attempt to control the content and focus of the items across formats. In other studies, there is no attempt to control for content and focus; instead, researchers take advantage of an existing test that happens to contain some multiple-choice and some constructed-response items. What we need are studies in which both multiple-choice and constructed-response items are developed in ways that allow each to "put their best foot forward," so to speak. To our knowledge, Fredericksen (1984) is one of the few researchers to consider the possibility that we may be introducing a source of bias when, for example, constructed-response items are generated by transformations from an existing set of multiple-choice items. He also is one of the few researchers to develop multiple-choice items from an existing set of constructed-response items. This study would extend the logic of his dual source approach to item generation. One could accomplish such goals with a procedure along the following lines:

- Identify a domain of interest, such as reading comprehension, response to literature, mathematical power, and so forth.
- Identify one group of item writers with reputations for developing first-rate multiple-choice items; identify a second group with equally strong reputations for constructed-response items.
- Set each group to work on developing a set of items for the domain of interest.
- When each group is finished, ask them to exchange item groups and, as best they can, transform each multiple-choice item into a constructed-response item and vice-versa.
- Create matched item sets, balanced for content and format.
- Administer to students, and evaluate relations between constructed-response and multiple-choice item subsets.

Garavaglia (2000) has completed just such a study in on NAEP mathematics items in the domain of algebraic representations of everyday experience. Garavaglia found that for the run-of-the-mill constructed response items that appear in NAEP, which tend to be little more than multiple-choice items in disguise, little advantage is gained by including the constructed response format. It was only when genuine performance tasks (e.g., multiple-step, problem-solving tasks with opportunities to write about one's work) were included in the mix that one can show the value added of constructed response items. It would be interesting to transfer this methodology to reading comprehension assessment.

The Role of Passage Difficulty in Reading Assessment. In the previous section, we outlined the dimensions of this issue, but here we deal with the interaction of passage difficulty and item format. If more "accessible" blocks, comprised of easier passages, are created and if we are thoughtful about how we design and generate items across blocks, we can determine whether response format (multiple-choice versus constructed-response) or passage difficulty (or some unique combination) is responsible for the current low information yields of constructed-response items in tests like NAEP and many state tests. It might be, for example, that students have a lot more to say when it is relatively easy for them to read, digest, think about, and even critique the texts they encounter. It might also turn out that difficulty interacts with student achievement level in such a way that easy passages provide opportunities for low-achieving students to shine whereas hard passages provide just the challenge that high-achieving students need to get involved in the assessment.

Revisiting the Sentence Verification Task. The SVT has never been able to build much of a constituency, and we are not sure why. Perhaps it is because it is viewed as too cumbersome a process to use on a regular basis. Perhaps it is because it seems limited to application with shorter texts. Perhaps it is because it confounds memory with comprehension and because it puts readers in the odd circumstance of having to decide between veridical restatement (is this exactly what the text said?) and semantic equivalence (does this mean the same thing as the text said?) when deciding what to do with a paraphrase item. If these concerns can be overcome, perhaps SVT deserves another round of experimental trials to determine if it can provide teachers and schools with useful information about comprehension. Perhaps the most significant challenge to SVT is whether test developers can apply it to longer passages of the type currently used in NAEP and an increasing number of state assessments. A demonstration of its usefulness with longer passages would go a long way toward increasing its perceived utility for standardized assessments.

Mixed Model Assessments

Earlier we suggested that the current trend is toward mixed models of assessment, along the lines of NAEP and several state assessment initiatives, not to mention a few standardized tests (e.g., the SAT–9 has the capacity to mix in some constructed response items). Given the increasing popularity of this model, we need to study its conceptual, psychometric, and pragmatic characteristics very carefully. Among other things, we need to know the relative contributions of various components, using research tools along the lines of those outlined in the previous section on item format studies. However, other initiatives are also called for within this domain of inquiry.

Compensatory Approaches to Achieve Particular Cut Scores. The New Standards Reference Exam (1998), currently marketed by HBJ, uses an interesting strategy to determine whether students meet a standard. The test is built on a mixed model, a combination of constructed responses to challenging passages and some very traditional multiple-choice responses to short passages, the stuff of which typical standardized tests are made. Test developers ask experts to determine the various combinations of scores on the multiple-choice and constructed-response portions of the test that would represent mastery of the construct. This is reminiscent of the admissions indexes used by universities to admit freshmen: different combinations of high school grade point average and SAT scores will get a student over the "admit" line—a high score on one component can compensate for a low score on the other. This raises an interesting possibility for creating comprehension scores to determine whether a particular standard (usually a cut score) has been met. Essentially this procedure caters to individual differences in item format preferences. The research question of interest is whether different combinations of scores on the two exams really can and do provide equal estimates of comprehension ability or achievement. For example, if we compared students who had achieved a common cut score with widely different patterns of reliance on constructed-response versus multiple-choice items, would we find that they both could meet a common external criterion, such as successful participation in a particular curriculum or activity (e.g., classroom discussion) or a score on a completely independent measure of comprehension.

Other (Not Readily Classifiable) Initiatives

Genre and Comprehension. One of the great mysteries of reading assessment is the persistent gap between performance on narrative texts and informa-

tional texts. In study after study, state assessment after state assessment, students consistently achieve higher raw scores on narrative texts. The question is why? Is the difference a function of opportunity to learn (we know that elementary students are exposed to at least 10 times more narrative than expository text)? Is it due to prior knowledge (other things being equal, do students know more about the everyday experiences depicted in stories than they do the propositional knowledge in expositions)? Or is there something about the way that we assess understanding of the two genres that creates an artifactual difference (maybe we test more central ideas in stories than in nonfiction)? A close content examination of item types, focus, and relation to the texts from which they are derived, followed by some small-scale experiments, seems appropriate to determine the source of this persistent finding.

Interest, Knowledge, Comprehension and the Idea of a Level Playing Field. We know that both interest and prior knowledge (which are themselves conflated) influence comprehension, but we have not really considered the consequences of these relations for assessment, at least not in recent years. Given what we know about the lack of generalizability of performance tasks and the capacity of passage effects to overwhelm cognitive process effects, we have an ethical obligation to get inside the quagmire that resides at the intersection of interest, knowledge, and comprehension. We know that our estimate of a given student's comprehension is dependent on the passages read. Our traditional solution to the influence of topical knowledge has been to make sure that we provide students with a wide range of short passages on widely variant topics. This practice more or less guarantees that students who have high general verbal ability will do best on standardized tests (see Johnston, 1984a, 1984b). And because all students have read the same passages, we seduce ourselves into believing that we have satisfied the fairness (level playing field) criterion. Perhaps we need to consider other metaphors for fairness. What if every student reading passages for which he or she possessed the maximum level of interest and knowledge, rather than every student reading the same passages, were considered to be the default fairness criterion? In other words, what might happen if we replaced the "level playing field" with "playing to readers' strengths" as a criterion of fairness? Yet, how would we know if we are indeed capitalizing on readers' strengths? It would be useful to examine performance (perhaps something like reaching a particular performance standard) as a function of variations in student interest and knowledge, where some sort of common task could be applied to a wide range of passages. Retelling could be applied across a wide array of passages. Another possibility is a core set of common, generic, constructed-re-

sponse questions; just such a set was used in 1994 NAEP special studies (as cited in Salinger & Campbell, 1998).

Examining the Consequential Validity of the Lexile Scale Framework. The Lexile scale (and the parallel DRP) holds great promise in helping teachers make difficult book placement decisions without the arduous effort of administering cumbersome benchmark book assessments or informal reading inventories. However, to our knowledge, the most important validity studies, especially for a measure that purports to impact practice, have not been conducted—examining the consequential validity of the book placements suggested by the analysis. Several studies are possible here. First, it would be interesting to compare the placement recommendations of teachers who possess different levels of experience and knowledge of children's literature with those provided by the Lexile process. Does the Lexile scale conform to the recommendations of more experienced and more knowledgeable teachers? Second, it would be useful to compare the experiences and understanding of students who read Lexile recommended books. It is one thing to make the connections to books through a common scaling procedure; it is quite another to validate the match in terms of all the cognitive, affective, and aesthetic features of a quality reading experience. In other words, can kids really read, appreciate, and benefit from books recommended by the Lexile process? And are they, in reality, more capable of reading those books than books with higher Lexile ratings?

CONCLUSION

Reading comprehension assessment has been a significant landmark in the educational landscape for just over 80 years. Its history is a remarkable story, one characterized by cycles of great hope and expectation alternating with periods of disappointment and frustration. A disappointment general to scholars throughout its history has been our persistent inability to see comprehension as it happens, what we have referred to as the phenomenological "click" of comprehension. Instead, they have had to content themselves with "artifacts" and residual traces of the comprehension process—indirect indexes of its occurrence. Each of these indirect indexes carries with it a cost, one that can be measured by the inferential distance between the evidence and the phenomenon itself. Many of the advances in comprehension assessment have, at least in a virtual sense, narrowed the distance between evidence and the process, providing us with greater confidence in our measures.

Other hopes and disappointments have been particular to specific periods. Two examples stand out: (a) the great expectations built up around performance assessments in the early 1990s, followed by the disappointment at their failure to stand psychometric tests of generalizability and reliability, and (b) the short-lived exhilaration so prominent in the late 1980s, which held a promise that we might find assessments that would match the models of instruction built on the principles governing allegedly challenging constructivist curriculum. Although the disappointments and frustrations are real, there has also been genuine progress. That progress is probably best represented by NAEP and some of our other mixed model, wide-scale assessments

And, of course, there is still much more to learn about how to measure a phenomenon that is as elusive as it is important. We have tried to outline, in our suggestions for future research, some of the issues that merit our attention. It is our modest hope that this chapter will serve as a catalyst for both lively conversation and concentrated work to improve our capacity to assess what is assuredly most important about reading—our ability to marshal all of our resources to make sense of the texts we encounter.

REFERENCES

Allington, R. L. (1984). Oral reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 829–864). New York: Longman.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255–291). New York: Longman.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*, 61–70

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*, 1–42.

Baumann, J., Jones, L., & Seifert-Kessell, N. (1993). Using think alouds to enhance children's comprehension monitoring abilities. *The Reading Teacher, 47*, 184–193.

Betts, E. (1946). *Foundations of reading instruction*. New York: American Book.

Binet, A. (1895). Assessment in reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 147–182). New York: Longman.

Bloom, B. S. (1956). Taxonomy of educational objectives. *Handbook 1: Cognitive domain*. New York: McKay.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*.

Bloom, D., & Green, J. (1984). Directions in the sociolinguistic study of reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 394–421). New York: Longman.

Board of Education, City of Chicago. (1984). *Chicago mastery learning reading*. Watertown, MA: Mastery Education Corporation.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197–211.

Bormuth, J. R. (1966). Reading: A new approach. *Reading Research Quarterly, 1*, 79–132.
Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.
California Learning Assessment System. (1994). *Elementary performance assessments: Integrated English-language arts illustrative material*. Sacramento: California Department of Education.
California Department of Education. (1987). *English language arts framework*. Sacramento, CA: Author.
Campell, J. R. (1999). Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. Unpublished doctoral dissertation, Temple University, Philadelphia, PA.
Campell, J. R., Voelkl, K. E., & Donahue, P. L. (1998). *NAEP 1996 trends in academic progress: Achievement of U.S. students in science 1969 to 1996, mathematics , 1973 to 1996, reading, 1971 to 1996 and writing, 1984 to 1996* (NCES 97–985). Washington, DC: U.S. Department of Education.
Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723–732.
Chapman, J. C. (1924). *Chapman unspeeded reading-comprehension test*. Minneapolis, MN: Educational Test Bureau.
Courtis, S. A. (1914). Standard tests in English. *Elementary School Teacher, 14*, 374–392.
Davis, F. B. (1944). Fundamental factors of comprehension of reading. *Psychometrika, 9*, 185–197.
Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*, 499–545.
Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly, 7*, 628–678.
Derrik, C. (1953). *Three aspects of reading comprehension as measured by tests of different lengths* (Research Bulletin 53–8). Princeton, NJ: Educational Testing Service.
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.
Dewey, J. (1938). *Experience and education*. New York: Collier Books.
Durrell, D. D. (1937). *Durrell analysis of reading difficulty*. New York: Harcourt Brace.
Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
Farr, R., & Greene, B. G. (1992, February). *Using verbal and written think-alongs to assess metacognition in reading*. Paper presented at the 15th annual conference of the Eastern Education Research Association, Hilton Head, SC.
Feuerstein, R. R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performance*. Baltimore: University Park Press.
Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202.
Freeman, F. N. (1926). *Mental tests: Their history, principles and applications*. Chicago: Houghton Mifflin.
Gagné, R. M. (1965). *The conditions of learning*. New York: Holt, Rinehart & Winston.
Garavaglia, D. (in press). *The impact of item format on depth of cognitive engagement*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Gates, A. I. (1937). The measurement and evaluation of achievement in reading. In *The teaching of reading: A second report*. Bloomington, IL.: Public School Publishing.

Gearhart, M., Herman, J., Baker, E., & Whittaker, A. K. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment* (CSE Tech. Rep. No. 363). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Ginn & Company. (1982). *The Ginn reading program*. Lexington, MA: Author.

Ginsberg, H. (1997). *Entering the child's mind: The clinical interview of psychological research and practice*. New York: Cambridge University Press.

Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA: National Academy of Education.

Goodman, K. S. (1968). *The psycholinguistic nature of the reading process.* Detroit, MI: Wayne State University Press.

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly, 5*, 1.

Goodman, Y. M., & Burke, C. L. (1970). *Reading miscue inventory manual procedure for diagnosis and evaluation*. New York: Macmillan.

Gray, W. S. (1916). *Standardized oral reading paragraphs*. Indianapolis, IN: Bobbs Merrill.

Gray, W. S. (1917). *Studies of elementary school reading through standardized tests* (Supplemental Educational Monograph No. 1). Chicago: University of Chicago Press.

Harris, C. W. (1948). Measurement of comprehension in literature. *The School Review, 56*, 280–289, 332–342.

Hartman, D. K. (1995). Eight readers reading: The intertextual links of proficient readers reading multiple passages. *Reading Research Quarterly, 30*(3), 520–561.

Hively, W. (1974). Introduction to domain-reference testing. *Educational Technology, 14*(6), 5–10.

Huey, E. (1908). *The psychology and pedagogy of reading.* Cambridge, MA: MIT Press.

Hunt, L. C. (1957). Can we measure specific factors associated with reading comprehension? *Journal of Educational Research, 51*, 161–171.

Illinois Goal Assessment Program. (1991). *The Illinois reading assessment: Classroom connections*. Springfield: Illinois State Board of Education.

Irwin, P. A., & Mitchell, J. N. (1983). A procedure for assessing the richness of retellings. *Journal of Reading, 26*, 391–396.

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

Johnston, P. H. (1984a). Assessment in reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 147–182). New York: Longman.

Johnston, P. H. (1984b). *Reading comprehension assessment: A cognitive basis*. Newark, DE: International Reading Association.

Jones, L. V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25*(7), 15–22.

Kapinus, B., Collier, G. V., & Kruglanski, H. (1994). The Maryland school performance assessment program: A new wave of assessment. In S. Valencia, E. Hiebert, & P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities* (pp. 255–276). Newark, DE: International Reading Association.

Kelly, E. J. (1916). The Kansas silent reading tests. *Journal of Educational Psychology, 7*, 63–80.

Kelly, T. L. (1935). *Essential traits of mental life*. Cambridge, MA: Harvard University Press.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 36–394.

Langer, J. (1995). *Envisioning literature: Literary understanding and literature instruction*. New York: Teachers College Press.

Linn, R. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1–16.

Linn, R. (1999). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

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*, 33–45.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111–151.

Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child Development, 48*, 986–992.

Matthews, M. (1996). *Teaching to read*. Chicago: University of Chicago Press.

McNamara, T. P., Miller, D. L., & Bransford, J. D. (1991). Mental models and reading comprehension. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, Vol. 2* (pp. 490–511). New York: Longman.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North-Holland.

Meyer, B. J. F., & Rice, E. (1984). The structure of text. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *The handbook of reading research* (pp. 319–352). New York: Longman.

Monroe, W. S. (1918). *Monroe's standardized silent reading tests*. Bloomington, IL: Public School Publishing.

Morrow, L. M. (1988). Retelling stories as a diagnostic tool. In S. M. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.), *Reexamining reading diagnosis: New trends and procedures* (pp. 128–149). Newark, DE: International Reading Association.

National Assessment of Educational Progress (NAEP) Reading Consensus Project. (1992). *Reading framework for the 1992 national assessment of educational progress*. Washington, DC: U.S. Printing Office.

National Center for Education and the Economy. (n.d.). *New standards reference exams*. San Antonio, TX: Harcourt Educational Publishers.

Naylor, M. (1972). *Reading skill variability within and among fourth-grade, fifth-grad, and sixth-grade students attaining the same reading achievement score*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Olshavsky, J. E. (1976–1977). Reading as problem solving: An investigation of strategies. *Reading Research Quarterly, 12*, 654–674.

Olson, D. (1994). *The world on paper: The conceptual and cognitive implications of writing and reading*. New York: Cambridge University Press.

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*. New York: Academic.

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

Pearson, P. D. (2000). Reading in the 20th century. In T. Good (Ed.), *American education: Yesterday, today, and tomorrow. Yearbook of the National Society for the Study of Education* (pp. 152–208). Chicago: University of Chicago Press.

Pearson, P. D., Calfee, R., Walker-Webb, T., & Fleischer, S. (2002). *The role of performance assessment in large scale accountability systems: Lessons learned from the inside*. Washington, DC: Council of Chief State School Officers.

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

Pearson, P. D., Garavaglia, D., Danridge, J., Hamm, D., Lycke, K., Roberts, E., et al. (in press). *The impact of item format on the depth of students' cognitive engagement*. Washington, DC: American Institutes for Research.

Pearson, P. D., Greer, E. A., Commeyras, M., Stallman, A., Valencia, S. W., Krug, S. E., et al. (1990). *The validation of large scale reading assessment: Building tests for the twenty-first century*. Urbana, IL: University of Illinois, Center for the Study of Reading.

Pearson, P. D., & Johnson, D. D. (1978). *Teaching reading comprehension*. New York: Holt, Rinehart & Winston.

Pearson, P. D., Spalding, E., & Meyers, M. (1998). Literacy assessment in the New Standards Project. In M. Coles & R. Jenkins (Eds.), *Assessing Reading to Changing Practice in Classrooms* (pp. 54–97). London: Routledge.

Pearson, P. D., & Stephens, D. (1993). Learning about literacy: A 30-year journey. In C. J. Gordon, G. D. Labercane, & W. R. McEachern (Eds.), *Elementary reading: Process and practice* (pp. 4–18). Boston: Ginn.

Pressley, M., & Afflerbach, P. (1995). Verbal protocols of reading: The nature of constructively responsive reading. Hillsdale, NJ: Lawrence Erlbaum Associates.

RAND Reading Study Group (Catherine Snow Chair). (2001). *Reading for understanding: Towards an R&D program in reading comprehension*. Washington, DC: RAND.

Rankin, E. F. (1965). The cloze procedure: A survey of research. In E. Thurston & L. Hafner (Eds.), *Fourteenth yearbook of the National Reading Conference* (pp. 133–150). Clemson, SC: National Reading Conference.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (Part 2, p. 371). Washington, DC: National Academy Press.

Richards, I. A. (1929). *Practical criticism*. New York: Harcourt Brace.

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

Rosenblatt, L. M. (1978). *The reader, the text, the poem: The transactional theory of the literary work*. Carbondale: Southern Illinois University Press.

Royer, J. M. (1987). The sentence verification technique: A practical procedure for testing comprehension. *Journal of Reading, 30*, 14–22.

Royer, J. M., & Cunningham, D. J. (1981). On the theory and measurement of reading comprehension. *Contemporary Educational Psychology, 6*, 187–216.

Royer, J. M., & Hambleton, R. K. (1983). *Normative study of 50 reading comprehension passages that use the sentence verification technique*. Unpublished manuscript, University of Massachusetts at Amherst.

Royer, J. M., Hastings, N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension tests. *Journal of Reading Behavior, 11*, 355–363.

Royer, J. M., Lynch, D. J., Hambleton, R. K., & Bulgarelli, C. (1984). Using the sentence verification technique to assess the comprehension of technical text as a function of subject matter expertise. *American Educational Research Journal, 21*, 839–869.

Royer, J. M., Kulhavy, R. W., Lee, J. B., & Peterson, S. E. (1986). The sentence verification technique as a measure of listening and reading comprehension. *Educational and Psychological Research, 6*, 299–314.

Rumelhart, D. E. (1977). Understanding and summarizing brief stories. In D. LaBerge & J. Samuels (Eds.), *Basic processes in reading perception and comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E. (1981). Schemata: The building blocks of cognition. In J. T. Guthrie (Ed.), *Comprehension in teaching* (pp. 3–26). Newark, DE: International Reading Association.

Salinger, T., & Campbell, J. (1998). The national assessment of reading in the USA. In C. Harrison & T. Salinger (Eds.), *Assessing reading: Theory and practice* (pp. 96–109). London: Routledge.

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House, 72*, 97–105.

Schreiner, R. L., Heironymus, A. N., & Forsyth, R. (1969). Differential measurement of reading abilities at the elementary school level. *Reading Research Quarterly, 5*, 1.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229–255.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Silver Burdett & Ginn. (1989). *World of reading*. Morristown, NJ: Author.

Smagorinsky, P. (2001). If meaning is constructed, what's it made from? Toward a cultural theory of reading. *Review of Educational Research, 71*(1), 133–169.

Smith, N. B. (1986). *American reading instruction*. Newark, DE: International Reading Association.

Spearitt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8*, 92–111.

Spiro, R., & Jehng, J. (1990). Cognitive flexibility and hypertext: Theory and technology for the linear and nonlinear multidimensional traversal of complex subject matter. In D. Nix & R. Spiro (Eds.), *Cognition, education, and multimedia: Exploring ideas in high technology* (pp. 163–205). Hillsdale, NJ: Lawrence Erlbaum Associates.

Starch, D. (1915). The measurement of efficiency in reading. *Journal of Educational Psychology, 6*, 1–24.

Stein, N. L., & Glenn, C. G. (1977). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *Discourse processing: Multidisciplinary perspective* (pp. 53–120). Norwood, NJ: Ablex.

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

Stenner, A. J., Smith, D. R., Horabin, I., & Smith, M. (1987). *Fit of the Lexlie Theory to item difficulties on fourteen standardized reading comprehension tests*. Durham, NC: MetaMetrics Inc.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 9*, 206–223.

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*, 323–332.

Thurstone, L. L. (n.d.). *Psychological examination* (Test 4). Stoelting.
Touchstone Applied Science Associates. (1995). *Degrees of reading power*. Benbrook, TX: Author.
Touton, F. C., & Berry, B. T. (1931). Reading comprehension at the junior college level. *California Quarterly of Secondary Education, 6*, 245–251.
Tuinman, J. J. (1974). Determining the passage-dependency of comprehension questions in 5 major tests. *Reading Research Quarterly, 9*, 207–223.
Tuinman, J. J. (1978). Criterion referenced measurement in a norm referenced context. In J. Samuels (Ed.), *What research has to say about reading instruction* (pp. 165–173). Newark, DE: International Reading Association.
Turner, A., & Greene, E. (1977). *The construction of a propositional text base* (Tech. Rep. No. 63). Boulder: University of Colorado Press.
Valencia, S., & Pearson, P. D. (1987a). *New models for reading assessment* (Reading Education Rep. No. 71). Urbana: University of Illinois Press, Center for the Study of Reading.
Valencia, S., & Pearson, P. D. (1987b). Reading assessment: Time for a change. *The Reading Teacher, 40*, 726–733.
Valencia, S., Pearson, P. D., Peters, C. W., & Wixson K. K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership, 47*, 57–63.
Valencia, S. V., Pearson, P. D., Reeve, R., Shanahan, T., Croll, V., Foertsch, D., et al. (1986). *Illinois assessment of educational progress: Reading*. Springfield: Illinois State Board of Education.
Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.
White, E. B. (1952). *Charlotte's web*. New York: Harper & Row.
Winograd, P., & Johnston, P. (1980). *Comprehension monitoring and the error detection paradigm* (Tech. Rep. No. 153). Urbana: University of Illinois Press, Center for the Study of Reading.
Yepes-Bayara, M. (1996, April). *A cognitive study based on the National Assessment of Educational Progress (NAEP) science assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.