The Assessment of Thoughtful Literacy in NAEP: Why the States Aren't Measuring Up

Anthony J. Applegate, Mary DeKonty Applegate, Catherine M. McGeehan, Catherine M. Pinto, Ailing Kong

Not all tests of reading are created equal, and the NAEP might distinguish reader levels more effectively than state tests.

The widespread publicity in the United States surrounding the "reading wars" that date back to the 1960s served to obscure one clear fact: When it comes to a definition of the nature of mature reading (the ultimate goal of all reading instruction), there is a remarkable level of agreement. This agreement exists among proponents of opposing philosophical camps; among reading theorists from the 19th, 20th, and 21st centuries; and among assessment specialists charged with measuring reading achievement at state, national, and international levels. The essence of the agreement is this: Mature reading involves thoughtful literacy—an ability to link the text with one's existing knowledge to arrive at a considered and logical response.

When Thorndike (1917) issued his oft-quoted comparison of reading to the act of human thinking, he had already been preceded in that line of reasoning by Huey (1908). Anderson (1984) cautioned his readers not to imagine that there is a simple, literal level of comprehension that does not require the reader to access a schema from world experience. When Goodman, Watson, and Burke (1996) described reading as an intellectually active process of creating meaning based on one's own experiences, Chall (1996) concurred, claiming that at all stages of development, reading depends upon full engagement with the text—its content, ideas, and values (p. 12). Even the National Reading Panel (National Institute of Child Health and Human Development, 2000) weighed in with their claim that comprehension requires that readers use knowledge of the world to make meaning of the text. In short, nowhere in the literature could we find a theorist or practitioner who would define mature reading as the ability to reproduce the message encoded in the text without also responding thoughtfully to it.

Our examination of all 50 U.S. state instructional frameworks and the specifications upon which the state assessments are based was equally unanimous and unequivocal. Specifications ranged from the "deep discussion and questioning" required in Alabama (Alabama Reading Initiative, 2001) to the ability "to use comprehension strategies to enhance understanding, to make predictions, and to respond to literature" in Tennessee (Tennessee Department of Education, 2007). No state defined reading solely as the ability to extract information from text. All state assessments expect at least some level of thoughtful response on the part of the reader.

It should come as no surprise that the National Assessment of Educational Progress (NAEP) Framework focuses on similar dimensions of reading. The 2007 NAEP Framework defined reading as including the ability

> to develop a more complete understanding of what is read, to connect information in the text with knowledge and experience, and to examine content by critically evaluating, comparing and contrasting, and understanding the effect of such features as irony, humor, and organization. (National Assessment Governing Board, 2006)

At the international level, the Progress in International Reading Literacy Study (PIRLS) assesses the ability

The Reading Teacher, 62(5), pp. 372–381
DOI:10.1598/RT.62.5.1

ISSN: 0034-0561 print / 1936-2714 online

to make inferences about ideas not explicitly stated, to interpret and integrate ideas, and to examine and evaluate content, language, and textual elements (Mullis, Kennedy, Martin, & Sainsbury, 2006).

The Ideal and the Real

The scope and unanimity of this agreement on the nature of mature reading offers reading educators an unprecedented opportunity to gear our instruction to the achievement of these goals. But a wide variety of researchers have found anything but a united front on how we approach the development of thoughtful literacy. For the most part they have observed classrooms that do not engage readers in thinking and responding to a text, but rather in memorizing and reciting its details (Allington, 2001; Brown, 1991; Elmore, Peterson, & McCarthey, 1996; Knapp, 1995; Tharp & Gallimore, 1989). What these researchers have observed seems at first glance to be nothing more than an instance of the educational community saying one thing and doing something completely different—a thorough disconnect between the ideal and the real.

One possible explanation for this seeming disconnect in the nation's literacy classrooms was put forth by Black and Wiliam (1998), who suggested that many teachers emphasize literal recall because they assume that they are preparing their students to perform well on accountability measures. This is an intriguing hypothesis and in fact state reading test results seem to support the thinking of the teachers. Well-publicized reports of assessment data suggest that a large proportion of students in a great number of states have achieved reading proficiency. The problem is that results from NAEP are not following suit. The state–NAEP comparisons for 2005 reveal that states reported a level of proficiency at a startling average rate of 40% higher than that found on NAEP (Wallis & Steptoe, 2007). In the face of what appear to be inflated levels of achievement on state tests, it is tempting to simply conclude that the state tests have "lowered the bar" in the face of demands stemming from the No Child Left Behind Act (Thomas B. Fordham Foundation, 2005).

We wondered whether the differences between state tests and NAEP ran deeper. The assumption that reading tests are roughly equivalent because they ask a reader to respond to questions about text is one that deserves closer examination. Furthermore, the issue of NAEP–state test equivalency has profound implications for the way that the U.S. educational community interprets its sometimes conflicting assessment data and uses those data as guides for future instruction. We set out to see if there were qualitative differences between state tests and NAEP in the assessment of thoughtful response.

Methods

We began by obtaining a sample of state achievement tests in reading comprehension. We elected to focus on fourth-grade assessments because NAEP is administered to both fourth- and eighth-grade samples. We obtained sample reading comprehension tests from the NAEP website and from the 20 most heavily populated states in the United States. We used the following four criteria to guide our state test selection process:

1. Fourth-grade sample tests were available online and included enough items to allow for reliable analysis.
2. These tests were specifically offered as samples designed to familiarize educators with the format and item types used to measure comprehension.
3. Items were accompanied by the passages upon which they were based.
4. Items were accompanied by designations of the level of thinking the items were intended to assess.

Sample tests from the following states met our screening criteria: California, Florida, Wisconsin, Illinois, New York, North Carolina, Pennsylvania, and Texas. Sources for the sample tests are found in Table 1. The average difference between results on the selected state assessments and NAEP was 40 points, exactly the average for all 50 states.

Analysis of Test Items

We set out to classify each item in our sample of tests according to three criteria, which

Table 1
Sample Test Items Retrieved From Websites

Test and date	Website
California (2007)	www.cde.ca.gov/ta/tg/sr/documents/RTqGr4ela.pdf
Florida (2001, 2007)	FCAT.fldoe.org/pdf/sample/0607/reading/FL07_STM_G4R_TB_cwf001.pdf fcat.fldoe.org/pdf/fc4rib0a.pdf
Illinois (2008)	www.isbe.state.il.us/assessment/htmls/sample_books.htm
National Assessment of Educational Progress (2007)	nces.ed.gov/nationsreportcard/ITMRLS/Startsearch.asp
New York (2007)	www.nysedregents.org/testing/elaei/07exams/home.htm
North Carolina (2005)	www.ncpublicschools.org/accountability/testing/eog/sampleitems/reading
Pennsylvania (2006–2007)	www.pde.state.pa.us/a_and_t/lib/a_and_t/2006-2007gr4ReadingItemSampler.pdf
Texas (2006)	www.tea.state.tx.us/student.assessment/resources/release/taks/2006/gr4taks.pdf
Wisconsin (2005)	dpi.state.wi.us/oea/readingptri.html

we selected because they might serve as possible explanations of the state–NAEP discrepancy:

1. Item type—Did the test item use an open-ended or multiple-choice format?
2. Item objective—Was the item intended to assess vocabulary knowledge, familiarity with genre, text organization, characterization, or text detail? The rubric we used for this classification is included in Figure 1.
3. Item purpose and cognitive demand—Did the item require the reader to understand the content of the text (text emphasis), or did the item require the reader to interpret the meaning of the text (higher order)?

Using several tests not included in the final sample, we met and discussed our classifications until we had achieved a solid level of confidence in our command of classification criteria. Each of us then independently classified each item in each sample test. The majority opinion was regarded as the final classification and we agreed on 96.1% of the Item objective and 94.7% of the Cognitive demand classifications, suggesting that the criteria we had developed could be used with a high level of confidence. The Item type criterion is self-explanatory (multiple choice or open ended), but Item objective and Item purpose/cognitive demand will require some elucidation.

Item Objective

Vocabulary Knowledge. To assess vocabulary, the test identifies a word or phrase (either underlining it in the text and referring the reader to its location, or quoting the sentence in which it appears). The intended objective for the test taker is to use the context clues available in the text to determine the meaning of the word or phrase and to select the best synonym or definition from among the choices presented.

Familiarity With Genre. These items may ask the reader to identify the particular genre in which a text selection is written. An effective genre item challenges readers to call to mind an entire range of ideas surrounding a particular kind of writing and to use those ideas to construct a framework in which the details in the text will unfold. This knowledge of the overall framework of the story can establish a set of expectations that will aid in an active response to text and a deeper level of comprehension. A more limited type of item asks readers to identify a specific convention of writing, such as a metaphor or simile, or to label a statement as fact or opinion.

Text Organization. Comprehension assessment items require readers to recognize the ways in which writers organize text or present their ideas. They may challenge readers to consider a writer's intent, distinguish between main and subordinate ideas, devise

Figure 1
Item Objective Classification Guidelines

Vocabulary items

- The reader must identify a word's meaning, presumably based on context clues provided in the text.
- The reader must identify the meaning of a phrase or figure of speech.

Genre items

- The reader must apply the definition of a genre type to identify the genre in which a piece of text is written.
- The reader is asked to identify a specific convention of writing, such as
  - A metaphor or simile
  - A fact or opinion
  - The difference between fantasy and reality
  - Writing techniques such as onomatopoeia, italics, parentheses, etc.

Organization items

- The reader is asked to
  - Detect the writer's purpose for writing a piece of text
  - Create or select an alternate title for a piece of text
  - Predict what is likely to happen, based on events that have already occurred
  - Identify an idea or ideas that are most important in the passage
  - Select or create a statement of the main idea or main event of a passage
  - Describe the way that the author choose or order the events or information in a passage
  - Describe an alternative ending for a story
  - Identify appropriate items that would fit into a schema map of text

Characterization items

- The reader must identify personality characteristics that are supported or developed in the text.
- The reader is asked to
  - Choose a word that best describes a character
  - Choose a word that best describes a character's feelings
  - Identify factors that may have motivated a character to act or to arrive at a set of beliefs
  - Identify characteristics that can be compared or contrasted to those of another character
  - Predict the action that a character would be likely to take based on what the reader has found out about that character
  - Identify a change in a character's behavior or attitude

Detail items

- The reader must recognize elements of the text that are stated verbatim in text or paraphrased.
- The reader must identify similarities or differences between selections or characters based on clearly stated elements of the text.

appropriate titles that do justice to the content of text, or to consider ways in which writers have framed arguments. Good text organization items assess a reader's appreciation for the variety of ways in which events and arguments can unfold and the levels of effectiveness associated with that variety.

Characterization. These items ask the reader to identify personality characteristics that are supported or developed in the text. They may require the reader to select a word that best describes a character or a character's feelings. They may ask that the reader identify a character's motivation or predict the actions that a character may take, based on what the reader has already learned about that character. Characterization items are not limited to narrative text; expository text also includes people whose characters are developed in the text. Effective characterization items call for a thoughtful analysis of human nature.

Detail. Detail items are geared toward the recognition of information stated directly, or nearly so, in text. They may involve comparison or contrast based on factual information. They may also involve recognition of the same information after it has undergone slight, extensive, or subtle paraphrasing. Effective detail items force the reader to turn

attention to significant text elements related to the central message of the text. However, weak items may direct a reader's attention to obscure facts and can, over time, distort a child's view of the nature of reading by encouraging memorization of less salient information.

Item Purpose and Cognitive Demand

Text emphasis items. We have defined Text emphasis items as those with answers stated verbatim in the text or so nearly so that they require only translation from one set of words to another (Applegate, Quinn, & Applegate, 2002). Pure verbatim items at the fourth-grade level are relatively rare. After all, the objective of test items is to discriminate between capable and less capable readers, and items that require readers to simply look up answers in the target text are unlikely to deliver that level of discrimination. Items that require the student to recognize the same message in a different linguistic form are far more likely to distinguish between capable and less capable readers, even if they require little thoughtful response.

But there are other variations of Text emphasis items to be considered in any analysis of assessments. We viewed as Text emphasis any item that used distractors so improbable that recognition of the correct answer required only a very low level of understanding of text. In a similar vein, we classified as Text emphasis those items that could be answered on the basis of test-taking skills without the need to thoughtfully respond to the author's message.

Higher Order Interpretation Items. This category of items is most reflective of and congruent with the definitions of mature reading comprehension discussed at the beginning of this article. These items require readers to draw logical conclusions based upon their understanding of the text and their own, often unique, related personal experiences. They may call upon readers to take and defend a stand, using elements from both text and personal experience. Higher order questions are often characterized by challenges to readers to compare or contrast characters, situations, conclusions, or even elements of personal experience that might be related to their understanding of text. These items constitute what we have referred to as *thoughtful literacy*, a response to text that reflects the ability of readers to use their life experiences to flesh out and make sense of the stories and information in text. Throughout our analysis, we took the position that even if a single distractor required readers to engage in a thoughtful interpretation of the text, it was enough to mark that item as requiring higher order thinking.

Results

Item Types and Item Objectives

The first point of comparison we examined between state tests and NAEP was the proportion of items that assessed comprehension in an open-ended as opposed to a multiple choice format (see Table 2). Our sample of comprehension items from NAEP included 57.0% open-ended items as opposed to an average of only 7.0% from our sample of state assessments. Among the state tests that we examined, only Florida made significant use of open-ended questions, but even so, their sample included less than half of the proportion of open-ended items that we observed in NAEP. Thus it appears that NAEP places a great deal more emphasis upon a reader's ability to construct and explain a response to text, whereas the state tests place a higher premium upon a reader's ability to recognize a response and distinguish it from other less adequate responses.

The results of our analysis of item objectives are presented in Table 3. Vocabulary items were seldom used in NAEP; only 1 of a sample of 62 items assessed vocabulary. In our state sample, however, two tests (California and Wisconsin) allocated more than 25.0% of their comprehension items to the assessment of vocabulary. On the whole, the state sample averaged over 17.0% vocabulary items. Our examination of vocabulary items in the state tests revealed one major difficulty—the inability of the test constructors to ensure that the target word is unknown to the reader. If the word is already known, the need to use context clues is short-circuited and the reader need only find the synonym from among the listed choices. Under those circumstances, the reader need not even to have read the passage, let alone comprehended it. In any case, it appears that NAEP de-emphasizes vocabulary items and our sample of state tests uses them regularly in comprehension assessment.

Genre items accounted for only 2 of 62 items on our NAEP sample in contrast to the state tests that on average allocated 11% of their items to the ability to identify elements of genre. The proportion of

Table 2
Percentages of Items in Item Type Categories on NAEP and Selected State Tests

Test	Multiple-choice	Open-ended
NAEP (*N* = 62)	43.0	57.0
California STAR (*N* = 36)	100.0	0.0
Florida FCAT (*N* = 32)	73.0	27.0
Illinois ISAT (*N* = 19)	95.0	5.0
Wisconsin WKSE (*N* = 21)	95.0	5.0
New York (*N* = 35)	88.0	12.0
North Carolina (*N* = 31)	100.0	0.0
Pennsylvania PSSA (*N* = 26)	92.0	8.0
Texas TAKS (*N* = 40)	100.0	0.0
State average	93.0	7.0

Table 3
Percentages of Items in Item Objective Categories on NAEP and Selected State Tests

Test	Vocabulary	Genre	Organization	Characterization	Detail
NAEP	1.6	3.2	25.0	46.0	24.0
California STAR	28.0	17.0	25.0	11.0	19.0
Florida FCAT	13.0	13.0	13.0	25.0	36.0
Illinois ISAT	16.0	16.0	32.0	32.0	4.0
Wisconsin WKCE	28.0	5.0	38.0	10.0	19.0
New York	14.0	9.0	31.0	26.0	20.0
North Carolina	20.0	16.0	10.0	35.0	19.0
Pennsylvania PSSA	15.0	12.0	35.0	23.0	15.0
Texas TAKS	18.0	0.0	35.0	33.0	14.0
State average	17.1	11.0	27.4	24.4	18.3

genre items in Texas and Wisconsin was very similar to NAEP's but California, Illinois, and North Carolina used more than 15.0% of their items to assess knowledge of genre. Weak genre items ask only that the reader apply a rote definition of a genre element to a piece of text or identify a convention of writing. Consequently, they can frequently be answered without reference to the text itself. For example, an item that presents readers with four sentences and asks them to identify which is an opinion is simply asking them to apply a definition and identify the sentence that cannot be proven. When we examined the 25 genre items that were included in our state test

sample, we found that 88.0% called for rote recall of definitions of textual elements.

Elements of text organization accounted for 25.0% of NAEP's items, and the state average was very similar. States that emphasized text structure and organization significantly less included Florida and North Carolina. Several of the states devoted more than a third of their items to this objective. Effective text organization items require readers to use text information to predict events or to hypothesize about alternative endings. However, some weaker items ask only that the reader identify the event that happened first or last in the text.

NAEP devoted a significantly higher proportion of its items to characterization (46%) than did the average state in our sample (24.3%) with California and Wisconsin weighing in with the fewest items devoted to analysis of characters. Even the state test with the highest proportion of characterization items (North Carolina) trailed NAEP by 11.0%.

Detail items accounted for 24.0% of the questions in the NAEP sample as compared with an average of 18.3% in the state sample. However, Illinois devoted significantly fewer items to detail than did NAEP; Florida devoted significantly more. The weakest of detail items call for the reader to exercise test-taking skills and eliminate distractors that are highly unlikely.

Recognition vs. Interpretation

We set out to determine whether state test developers primarily intended to measure the reader's ability to recognize information or to interpret that information. The item designations proposed by the state tests and NAEP are presented in Table 4 under the columns labeled Intended. We then assessed what we judged as the actual cognitive demands of those same items, and those results are summarized in the adjoining columns under the label Actual.

The data in Table 4 suggest that the state and NAEP test developers place a great deal of emphasis upon the interpretation of the text. With the exception of California and Texas, each of the state tests, as well as NAEP, report that more than half of their comprehension test items assess the reader's ability to think about text and to draw conclusions that go beyond mere memory for details. This proportion of items suggests that the intent of nearly all of the tests is congruent with the universally accepted definition of the mature reader we discussed earlier.

But an examination of our actual item demand classifications reveals that NAEP called for higher

Table 4
Comparison of Percentages of Items in Intended and Actual Cognitive Demand Categories on NAEP and Selected State Tests

Test	Text emphasis		Higher order	
	Intended	Actual	Intended	Actual
NAEP	6.5	32.2	93.5	67.8
California STAR	55.6	91.7	44.4	8.3
Florida FCAT	40.6	84.4	59.4	15.6
Illinois ISAT	45.5	78.9	54.5	21.1
Wisconsin WKCE	40.0	85.7	60.0	14.3
New York	40.6	71.4	59.4	28.6
North Carolina	35.5	71.0	64.5	29.0
Pennsylvania	32.5	69.2	67.5	30.8
Texas TAKS	50.0	67.5	50.0	32.5
State average	42.5	77.5	57.5	22.5

order interpretation more than twice as frequently as the highest ranked state test (Texas), more than three times as frequently as the average state test, and more than eight times as frequently as the lowest ranked state test (California). A breakdown of various item objectives can shed some further light on the differences between our sample of state tests and NAEP.

Text organization items and characterization items were much more challenging on state tests, weighing in at 43.0% and 40.0% Higher order, respectively. Thus it seems that these two item classes have the greatest potential to elicit thoughtful responses from readers. However, NAEP organization items were 93% Higher order and characterization items were ranked at 79% Higher order. In both cases, the challenge to think on NAEP was roughly double that on our sample of state tests.

Open-ended items seem to have a great deal more potential to assess reading as a linking of text with an individual's unique experiences than forced choice items. What we found surprising was that fully one half of the open-ended items on our state sample fell into the Text-emphasis category; only 15% of NAEP's open-ended items assessed pure recognition of text elements.

The Analysis of Cognitive Demand

We found that test developers tended to classify multiple-choice items based on the question stem, often without regard to the quality of the distractors. It is the nature of a multiple-choice item, however, to require the reader to select the best choice and eliminate the incorrect ones. For example, one state test included a story about a captured sparrow kept in a rattan cage and sold as a pet to several different characters. Each time he changes hands, the sparrow begs his new owner to set him free but he cannot be understood. Finally a worker buys the sparrow and takes him home to cheer his daughter who is confined to bed with an illness. The girl immediately understands the sparrow's predicament and sets him loose, asking him to fly in freedom for both of them.

One multiple-choice test item asks the reader to determine how the bird and the little girl are alike, an item stem that cuts directly to the heart of the story and one that is clearly intended to assess thoughtful response. The correct response is that they both know what it is like to be trapped inside. The incorrect responses are that both the girl and the bird are very sick (directly contradicts passage content), both like rattan cages (illogical), and both are free to travel wherever they want (directly contradicts passage content). The nature of these distractors is such that only a reader who has understood very little of the gist of the story could possibly choose one of them. Thus a question that should require a great deal of thoughtful consideration requires in the final analysis only an understanding of the barest facts of the story.

Similarly, the consistent use of a particular type of distractor may convert even very thought-provoking items into exercises in test-taking skills that have little to do with actual comprehension. In many instances, test constructors overused a form of distractor labeled a Quiz Contestant response (Applegate, Quinn, & Applegate, 2006), in which the distractor provides a logical or sensible answer, but one that is drawn from pure experience without reference to the text. For example, a test item based on the passage described above asked why the characters could not understand what the bird was saying. One of the distractors was that they were not used to hearing bird sounds. If it were true, that fact would logically account for the inability of the characters to understand the bird, but there is no indication in the text that this was the case. If students are taught to identify and eliminate such distractors, they may circumvent any intended demands for thoughtful response. Consequently, we classified such items as Text emphasis.

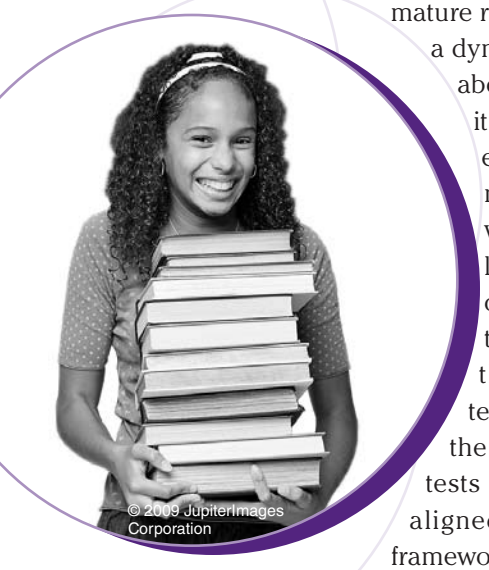
We must emphasize that multiple-choice test items do indeed have the potential to elicit thoughtful responses from readers; it is the content of the items that determines cognitive demand. A question that asks a reader to select which event fits in a sequence of events taken directly from the text differs significantly from an item that asks the reader to predict what is likely to happen next, based on the events in a story. By the same token, asking a reader what motivated a character to behave in a particular way when that motivation is stated directly in the passage is in no way equivalent to asking the reader to identify which character in the story would be most likely to agree with a statement.

Conclusions, Cautions, and Recommendations

While our sample of state tests was limited to eight, our study supports the conclusion that not all tests

labeled *reading comprehension* are measuring the same objectives. Our analysis suggests that there are qualitative differences between NAEP and our sample of state tests that may have contributed to the state–NAEP achievement gap. To assume that the state tests are simply an easier version of the same assessment seems to us to be a serious oversimplification.

In summary, we found that NAEP uses far more open-ended items in its assessment of reading, uses far fewer vocabulary and genre items, and demands far more thoughtful response than any of the state tests in our sample. Our analysis showed that just under one third of the items in NAEP centered on Text-emphasis reading and more that three quarters of the items in our state test sample fell into that category. Thus NAEP is far more congruent with the widely accepted definition of mature reading comprehension as a dynamic process of thinking about what we read and how it fits in with our experiences and values. NAEP is also much more closely aligned with the frameworks published by a vast majority of the states, frameworks that unanimously call for thoughtful responses to text. It is difficult to avoid the conclusion that the state tests are not particularly well aligned with their own testing frameworks. If that is the case, they may not be effective assessments of the results of their own curricula.



A significant advantage of NAEP is related to the complaint voiced by many educators—that assessment is driving curriculum, and teachers are being pressured to teach to the test rather than toward the achievement of a set of clearly articulated goals. But if that assessment tool is NAEP, a measure that seems to us to be effectively assessing widely agreed upon goals, then teaching to a test that assesses thoughtful response may actually work to the advantage of a great many of our children. If, on the other hand, we simply assume that all tests are equal determinants of achievement, that assumption may lead us to a national educational disaster.

What may be called for at this point is a national dialogue about the nature of our goals for reading instruction, and the means we must select to assess them. NAEP expects readers to be able to read, understand, and respond to text, but much of our instruction and assessment seems geared toward recognizing the literal content of the text (Durkin, 1978; Pressley et al., 2001). There is little doubt that U.S. students are better able to perform on tests that require remembering than on tests that require a thoughtful response to text (Donahue, Voelkl, Campbell, & Mazzeo, 1999). And once we recognize that fact, we need to decide if we as a nation are satisfied with that state of affairs. Our results suggest strongly that not all tests are created equal and that state–NAEP comparisons suggest that all is not well. To assume that the state test results are showing us that a sizeable majority of our children are on the road to mature reading is a potentially serious error, one that may have critical educational repercussions.

The state tests that we examined represented a string of missed opportunities to assess thoughtful response. For example, one test included a narrative about a pair of twin turtles who decide to play a trick on a hippopotamus by having one twin challenge the hippo to a swimming race. The hippo knows that he is much faster but when he arrives at the river bank, there is the other twin waiting for him. Rematches bring about the same result, and the hippo is forced to admit that the turtle is the faster swimmer. This is a simple narrative but it has multiple layers of meaning. The assessment of comprehension for this passage consists of four questions: two vocabulary items, an item that asks the reader to identify a hyperbole, and an item that asks the reader to select a word (tricky, lazy, brave, or stingy) to describe the turtles. In the case of the last item, it is difficult to avoid the conclusion that only a reader who had failed to understand the basic gist of the text could select one of the distractors.

It is important to note that if a test includes a sufficient number of such distractor items, it does not lose its value entirely. In most cases, it can serve as a discriminator between below basic readers on the one hand, and a combination of basic, proficient, and advanced readers on the other. However, when it is asked to discriminate among basic, proficient, and advanced readers, it simply does not include enough thought-provoking questions to accomplish the task. Among the tests that we analyzed, only NAEP required enough of a variety of thinking tasks

to discriminate among these groups. This observation alone may account for a great deal of the discrepancy between state and NAEP scores.

Finally, let us return to the hypothesis of Black and Wiliam (1998) who suggested that many teachers emphasize literal recall in their classrooms under the mistaken assumption that their students will perform well on accountability measures. Our analysis suggests that such literal-minded teachers may not be so mistaken after all, if we stay the course and continue to assess comprehension as if it consisted primarily of literal recall. But we run the risk of creating a growing number of students who perform well on state tests, yet continue to view reading as an exercise in literal recall of information, an exercise that does not require a spontaneous thoughtful response.

Our analysis of the content of state tests and NAEP suggests that teachers who encourage their students to engage thoughtfully with text and attend to the ways that details support thoughtful conclusions will prepare them to do well on both state and national accountability assessments. But as literacy professionals, we must call upon our state accountability tests to do much more to assess higher order interpretation of text if more of our children are ever to achieve the vision of mature reading that stands at the very core of the field of reading and literacy instruction.

References

Alabama Reading Initiative. (2001). *Comprehension strategies, grades K–1*. Retrieved November 30, 2007, from ftp://ftp.alsde.edu/documents/50/K-1_COMPREHENSION_2001.doc

Allington, R.L. (2001). *What really matters for struggling readers: Designing research-based programs*. New York: Addison-Wesley Higher Education.

Anderson, R.C. (1984). Role of the reader's schema in comprehension, learning and memory. In R.C. Anderson, J. Osborn, & R.J. Tierney (Eds.), *Learning to read in American schools: Basal readers and content texts* (pp. 243–257). Hillsdale, NJ: Erlbaum.

Applegate, M.D., Quinn, K.B., & Applegate, A.J. (2002). Levels of thinking required by comprehension questions in informal reading inventories. *The Reading Teacher, 56*(2), 174–180.

Applegate, M.D., Quinn, K.B., & Applegate, A.J. (2006). Profiles in comprehension. *The Reading Teacher, 60*(1), 48–57. doi:10.1598/RT.60.1.5

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

Brown, R.G. (1991). *Schools of thought: How the politics of literacy shape thinking in the classroom*. San Francisco: Jossey-Bass.

Chall, J.S. (1996). *Stages of reading development* (2nd ed.). Fort Worth, TX: Harcourt Brace.

Donahue, P.L., Voelkl, K.E., Campbell, J.R., & Mazzeo, J. (1999). *The NAEP 1998 reading report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

Durkin, D. (1978). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly*, *14*(4), 481–533.

Elmore, R.F., Peterson, P.L., & McCarthey, S.J. (1996). *Restructuring in the classroom: Teaching, learning, and school organization*. San Francisco: Jossey-Bass.

Goodman, Y.M., Watson, D.J., & Burke, C.L. (1996). *Reading strategies: Focus on comprehension* (2nd ed.). Katonah, NY: Richard C. Owens.

Huey, E.B. (1908). *The psychology and pedagogy of reading*. New York: Macmillan.

Knapp, M.S. (1995). *Teaching for meaning in high-poverty classrooms*. New York: Teachers College Press.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College, International Study Center.

National Assessment Governing Board. (2006). *Reading framework for the 2007 National Assessment of Educational Progress*. Retrieved December 18, 2007, from www.nagb.org/frameworks/reading_07.pdf

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Pressley, M., Wharton-McDonald, R., Allington, R.L., Block, C.C., Morrow, L., Tracey, D., et al. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading*, *5*(1), 35–58. doi:10.1207/S1532799XSSR0501_2

Tennessee Department of Education. (2007). *Academic standards*. Retrieved November 30, 2007, from www.state.tn.us/education/ci/english/grade_4.shtm/

Tharp, R.G., & Gallimore, R. (1989). Rousing schools to life. *American Educator, 13*(2), 20–25, 46–52.

Thomas B. Fordham Foundation. (2005, October 19). *Gains on state reading tests evaporate on 2005 NAEP* (Press release). Retrieved October 4, 2007, from www.edexcellence.net/detail/news.cfm?news_id=404&id=

Thorndike, E.L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology*, *8*(6), 323–332. doi:10.1037/h0075325

Wallis, C., & Steptoe, S. (2007, June 4). How to fix No Child Left Behind. *TIME, 169*, 34–41.

A. Applegate teaches at Holy Family University in Philadelphia, Pennsylvania, USA; e-mail Tapple1492@aol.com. M. Applegate teaches at St. Joseph's University; e-mail Mapple1492@aol.com. McGeehan teaches at Gwynedd-Mercy College in Gwynedd Valley, Pennsylvania, USA; e-mail catherinmcgeehan1@comcast.net. Pinto teaches for Gateway Regional High School District in Woodbury Heights, New Jersey, USA; e-mail catpin@msn.com. Kong teaches at St. Joseph's University in Philadelphia, Pennsylvania, USA; e-mail akong@sju.edu.