HOW AND HOW NOT TO PREPARE STUDENTS FOR THE NEW TESTS

Timothy Shanahan

The curriculum director regarded me with skeptical abhorrence. She herself had invited me into the school district with the idea that I would help raise reading achievement—or, more exactly, she thought that I would support what she'd been doing to raise reading scores, and she hoped that my approval would convince teachers to cooperate with her.

The problem was that I didn't know the script. I thought she was asking for my professional opinion, but she wanted affirmation. She laid out her logical test-score improvement plan and I respectfully disagreed. I wasn't trying to be difficult, but her path to test-score nirvana wasn't sensible, and I told her so.

What she wanted her teachers and principals to do has played out in thousands of schools around the country—without much to show for it except a lot of frustrated teachers and kids and reading scores that continue to languish.

"Data-driven school reform" is what I am talking about. This movement has swept the administrative ranks of schools with the idea that we can make reading instruction much more specific and intensive in ways that will raise test scores.

A perfectly reasonable way to use test data is to identify which students are struggling with reading, then provide them with additional instruction of some kind. But data-driven reforms often encourage educators to go further than this. They champion the idea that item analysis of standardized tests will allow teachers to know not only who may be having trouble but also which skills these students lack.

That would mean educators could use test data to figure out which standards are being learned and which are not. Then, teachers could give kids specific instruction in their areas of weakness—making them better readers and improving their test scores to boot.

Since 2010, 43 states have adopted the Common Core State Standards (CCSS) as the basis of reading instruction in their schools, and 33 of these states will be using these new innovative tests to evaluate the accomplishment of the CCSS standards. The Partnership for Assessment of Readiness for

Timothy Shanahan is a Distinguished Professor Emeritus at the University of Illinois at Chicago, Illinois, USA; e-mail shanahan@uic.edu.

"'Data-driven school reform'...has swept... schools with the idea that we can make reading instruction more specific in ways that will raise test scores."

College and Career (PARCC; www.parcconline.org) and Smarter Balanced Assessment Consortium (SBAC; www.smarterbalanced.org) tests are pioneering or exploiting many new test item features, including multipart items, multiple-choice items with multiple correct answers, technological responses (e.g., highlight parts of the text, drag and drop), and multiple texts with items requiring comparisons of the texts.

Already, new test prep materials are starting to appear on the Internet, and teachers ask me all the time how they can best prepare students to respond to these new items. Many educators assume that, since these tests will be aligned to the CCSS, they will detect which standards the students are meeting: Johnny meets Reading Standard 4 but is still not getting Standard 7.

The problem is that it doesn't work that way. It hasn't in the past, and it won't with these new tests, either.

The Problems With Item Analysis

Research long ago revealed an important fact about reading comprehension tests: they only measure a single factor (Davis, 1944; Spearritt, 1972). What I mean by that is that standardized comprehension tests do not measure multiple skills; they measure a single global one: reading comprehension. They don't reveal students' abilities to answer main idea questions, detail questions, inference questions, drawing conclusion questions, or anything else.

Why not?

Two reasons.

The first has to do with the nature of reading comprehension itself. Readers do not work their way through texts trying to apply a set of skills analogous to question types. Reading is more of a language activity. One has to interpret and interrelate a hierarchy of language features simultaneously to make sense of an author's message.

While students may have missed the main idea question, you cannot assume from this that the main idea part of their brains weren't working. There are just too many alternative explanations for the slip-up: (1) It was a long passage. Maybe these students thought it looked too hard, so they didn't read it. That means it was a confidence problem rather than a main idea one. (2) The students' reading speed was really low, so they just couldn't get to this item. (3) The students tried to read the passage, but with decoding skills so limited that an insufficient number of words were identified to allow a grasp of the main idea. (4) The students' decoding was great, but there were unknown vocabulary words. Can't get the main idea if you don't know what the words mean. (5) They recognized all the words and knew their meanings, but the key idea required to answer this question was embedded in a particularly complex sentence (26 words long, with two dependent clauses, in passive voice). Thus, the error was due to students' inability to untie the complex syntax. (6) The students could make sense of the sentences, but there were a lot of synonyms and pronouns, and keeping those all connected properly proved overwhelming—in other words, a cohesion problem.

ACT, the college testing people, analyzed their own tests and found that none of the question categories helped explain student performance (ACT, 2006). They tried combining their questions into various categories, but there simply was no consistent pattern in how students responded to various types of questions.

What *did* make a difference in comprehension performance? Text complexity. Students were less likely to answer questions correctly about challenging texts, and they answered more questions about the easier texts. That means, if the text is easy enough, students can answer any type of question, and if the text is complicated enough, they will struggle with even the supposedly easiest types of questions.

That means reading comprehension tests measure how well students read texts, not how well they execute particular reading skills (e.g., the question types).

Another reason item types don't discriminate in the way assumed by data-driven reformers has to do with how reading comprehension tests are designed. Reading tests need to be reliable (that is, the tests should render

"Standardized comprehension tests do not measure multiple skills; they measure...reading comprehension."

"Students may have missed the main idea question, but not because the main idea part of their brain wasn't working."

the same results on repeated administrations), and the test should be able to distinguish good and poor readers. To accomplish reliability, tests usually include 30 to 40 items. Test-makers also look for strong point-biserial correlations. That just means that they make sure that each item has a reasonably strong correlation to the overall results of the test—that is, each item helps to separate the good readers from the strugglers.

To accomplish this, test-makers usually try out more items than they need. Let's say your test is to include four passages with 10 questions each. You would probably start with 12 or 15 questions for each passage. That way, you could select the items with the best psychometric properties while dropping the rest. If an item were not highly correlated with the other items, it wouldn't be used. This approach serves to increase how similar the items are, which enhances reliability and validity, yet it also reduces the chances of there being any identifiable differences among the question types.

Test designers are satisfied by being able to determine how well students read and by arraying students along a valid reading comprehension scale. They purposely avoid making claims about the ability of their tests to determine how well students perform on the supposedly underlying skills represented by the question types. They know that the items collectively assess reading comprehension, but that separately—or in small sets of items aimed at particular kinds of information—the items can tell us nothing meaningful about how well students can read.

There are examples of reading comprehension test designs that have tried to provide more fine-grained information, but this has proven to be very difficult and it is not typical of reading comprehension tests generally. The National Adult Literacy Survey (NALS), for example, set out to develop a test that would result in three separate scores (Kirsch & Jungeblut, 1986). But even in that case, the three scores had to do with the nature of the texts rather than with the categories of questions. (NALS assesses how well adults read prose, how well they read documents—that is, graphic materials—and how well they handle the arithmetic operations embedded in some texts.) To make it possible to arrive at three reliable subscores, NALS had to be a much longer test than usual, and even with that, the three parts are moderately correlated with each other.

But What About the New PARCC and SBAC Tests?

Given how innovative these new PARCC and SBAC tests are, won't they be able to provide the kind of specific diagnostic information that past tests could not? In a word, no. These new tests won't be able to alter the nature of reading comprehension or the technical requirements for developing reliable test instruments.

If you have any doubts about this, go to the PARCC website and examine the nature of the claims made on the basis of student performance.

PARCC provides a series of sample items, including explanations of how the items align to the standards and evidence statements. I'll focus on a third-grade example (www.parcconline.org/sites/parcc/files/PARCCGrade3.pdf). This example includes an informational text and three test questions of different types.

The claim that PARCC makes for each item is identical: "Students read and demonstrate comprehension of grade-level complex informational texts." Its evidence that this is the case is based on the alignment of the questions with the standards. Since the standards define comprehension, the fact that the question reflects the standard proves that it is getting at comprehension. A little circular, maybe, but the key point is that PARCC is not claiming that performance on an item demonstrates how well a student "provides an explanation of how key details in a text support the main idea" or how well the student "provides explicit references to the text as the basis for the answers."

There is a good reason why neither PARCC nor SBAC has made any claims about how well its instruments will evaluate student performance on particular standards or skills. These consortia make no such claims because their instruments are not fine-grained enough to provide such information, thank goodness (think of how long these tests would need to be to provide such information!).

"Won't PARCC and SBAC provide specific diagnostic information? In a word, no."

Teaching to the Test

At this point, you might be having the same feeling as the curriculum director with whom I began. What good are these tests if we can't identify the specific skills students may lack?

Her plan was to give kids lots of practice answering test questions of particular types. The problem with such a plan is that it doesn't actually work. As you've seen, students may fail to answer particular questions, but not because they can't execute those skills. Giving students a lot of practice with those kinds of test items is not likely to improve achievement at all. It could even lower it, since there are better things that students could be doing to prepare for such testing.

The PARCC and SBAC tests may not provide the kinds of specific diagnostic information you may desire, but they should be able to offer some useful information. These tests will ask students to read extensive amounts of literary and informational text, to answer meaningful questions about these texts, and to provide explanations of their answers. These tests should do a pretty good job of showing how well students can read and comprehend challenging texts without teacher support.

Prepare students to excel on these tests not by focusing instruction on question types but by making students sophisticated and powerful readers. I encourage the following five steps:

1. *Have students read extensively within instruction. These tests measure reading ability, and you are not likely to develop reading ability without letting students read.*

It has been widely documented that students don't read much in school (Allington, 2011). The solution to this is, not a free-reading period, but including reading within your lessons. There is no excuse for having students read as little as they often do during reading comprehension lessons. Round-robin reading involves one child at a time in reading. Teachers like it because it provides control and it lets them observe how well a student is reading, but a reading comprehension lesson, except with the youngest children, should emphasize silent reading—and lots of it. Not only should students be reading within their reading class, but it should also be part of their social studies, science, and math lessons, too. Because this reading is done within lessons, teachers need to hold students accountable for gaining knowledge and understanding from what they are asked to read.

"Don't focus on question types...but on five steps that can make students sophisticated and powerful readers."

2. *Have students read increasing amounts of text without guidance or support. Performing on a test is like delivering a monologue, not like participating in a conversation.*

Often, lessons involve students in brief amounts of reading punctuated by class or group discussion. Students might read a paragraph or a page, followed by teacher questions. This model is not a bad one. It allows teachers to focus student attention on key parts of the text and to sustain attention throughout. However, the stopping points need to be progressively spread out over time. Perhaps early in the year, a teacher might have the group read a page at a time with follow-up discussion or activity. At some point, this reading regimen should be expanded to two or three pages' reading without interruption. The shortest prototype that PARCC or SBAC has released is a 550-word passage for third graders. It is essential that students gain extensive experience reading texts this long, and even longer, without teacher intervention or support. Increasing student stamina and independence in this way should be a goal of every reading teacher.

3. *Make sure the texts are rich in content and sufficiently challenging. Lots of reading of easy text will not adequately prepare students for dealing with difficult text.*

The CCSS established text levels that students should be able to read at grades 2–12, and PARCC and SBAC will assess reading with texts written at those challenge levels. In the past, elementary teachers have placed students in texts that matched their reading levels (Shanahan, 2013). But this is not the best way to enable students to handle more challenging text. Make sure the texts that you are assigning are sufficiently difficult, and provide students with scaffolding that allows them to perform well with these texts (Shanahan, Fisher, & Frey, 2012). This means providing fluency instruction with such texts and preteaching some of the key vocabulary words. It might require guidance with sentence grammar, text structure, or cohesion. In any event, it is essential that students learn to make sense of texts as difficult as those they will be expected to read on the tests.

4. *Have students explain their answers and provide text evidence supporting their claims.*

Studies suggest that students are not engaged in classroom activities with sufficient intellectual depth and that

involving them in such activities can have a positive impact on learning (Rowan & Correnti, 2009). One way that the CCSS emphasize intellectual depth is by requiring that students be able to use texts as the basis of their own arguments. Arguments require claims based upon reason and evidence, so involving students in such intellectual explorations of the texts they read will move them in the right direction. I would not expect such practices to enhance performance on any particular item types; however, I do believe that they will require students to read and reread texts in productive ways.

5. *Engage students in writing about text, not just in replying to multiple-choice questions.*

Most of the PARCC and SBAC items are multiple-choice. Nevertheless, research shows that writing about text enhances reading comprehension. Graham and Hebert (2010) in a meta-analysis of dozens of studies found that writing about text was a more powerful stimulant to learning than reading alone, reading and rereading, reading and discussing, or reading and studying. Although writing text summaries and syntheses may not look like the tests students are being prepared for, this kind of activity should provide the most powerful and productive kind of preparation.

In conclusion, the point here is a simple one: if you want your students to perform at their best on the new Common Core assessments, you will accomplish that not by having students practice items like those you will find on the PARCC and SBAC tests, but by teaching students to read.

REFERENCES

ACT. (2006). *Reading between the lines*. Iowa City, IA: American College Testing.

Allington, R.L. (2011). What at-risk readers need. *Educational Leadership, 68*(6), 40–45.

Davis, F.B. (1944). Fundamental factors in comprehension in reading. *Psychometrika, 9*(3), 185–197.

Graham, S., & Hebert, 2010. *Writing to read: Evidence for how writing can improve reading*. New York, NY: Carnegie Corporation.

Kirsch, I.S., & Jungeblut, A. (1986). *Literacy profiles of America's young adults*. Princeton, NJ: Educational Testing Service.

Rowan, B., & Correnti, R. (2009). Measuring reading instruction with teacher logs. *Educational Researcher, 38*(7), 549–551.

Shanahan, T., & Duffett, A. (2013). *Common Core in the schools: A first look at reading assignments*. Washington, DC: Thomas B. Fordham Institute.

Shanahan, T., Fisher, D., & Frey, N. (2012). The challenge of challenging text. *Educational Leadership, 69*(6), 58–63.

Spearritt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8*(1), 92–111.