

Summary of conversations with NY State Education staff on the 2006 4th grade writing test

Yesterday I spoke at length with two fairly forthcoming employees of the state Department of Education in Albany about some of the issues involved in this year's 4th Grade ELA test. Here's a summary of those conversations.

I'll give a few overall impressions and conclusions at the end, but here they are in a nutshell: The ELA test is not designed for, and should not be used for, making fine distinctions among students' performance – especially for high-stakes decision-making such as middle-school admissions, personnel evaluation, student promotions, etc. It's a state test, but the city uses it for all kinds of purposes not necessarily endorsed by the state.

I first spoke with **Dawn Thompson** of the Office of Reporting and Information Services, which has responsibility for getting the test results out. She explained that this year's ELA is very different from last year's ("apples and oranges"), because this year's test had to be re-done in order to comply with new federal requirements for testing at all grades from 3rd through 8th. The grading, scaling, and "cut scores" (deciding where to draw the lines between categories 1, 2, 3 and 4) were all made more complex and difficult by the fact that "this is a whole new type of test."

She explained that the state was also developing a new data collection system that would allow students to be tracked over years, adding more delay.

She said the individual reports will be sent to schools within 2 weeks. She also said that the school principals will have our students' actual tests, and that it is up to them whether to make them available for parents' inspection. (The next person I spoke with gave me a different impression: that the principals were *required* to allow parents to review their children's actual test, upon request.)

I asked how parents can appeal their children's scores. She said they can't, under a state Education Dept. policy that's been in place since these tests started. (Of course, just because there's a "policy" doesn't mean they don't do it.)

"What if the tests are determined to be flawed," I asked. "They have already been determined not to be flawed," she replied.

Ms. Thompson said she was not aware that the percentage of NYC fourth graders scoring a four on the 2006 test had declined by 60% compared with last year. She said that as far as she knew, there was no effort by the state to evaluate this decline. She stressed again that this year's test was "a completely different kind of test" than last year's.

Next I spoke with **Diane Knight**, who coordinates the design, development and administration of the tests with the Education Dept.'s Office of Assessment, who was quite helpful.

Here's how Ms. Knight said the tests are developed:

- * The state solicits bids from the small universe of companies in the world of test development. The ELA is a complex and expensive test (even in "regular" years when it's not being revamped to fit into a sequence of grade 3-8 annual tests), because: a) it's not a standardized test comparing children against one another, but rather a test to evaluate student performance against a set standard, prescribed in the state curriculum, and; b) the state insists on making each year's test public after it's been administered, so questions and materials can't be re-used.
- CTB/McGraw Hill has a multi-year contract to develop and help administer these tests, through 2008. How much? A "huge" amount. [Let's find out.]
- CTB starts the ELA test development process by identifying some written passages on which questions can be based. These are very difficult and expensive to find (for reasons I'm not exactly clear on) so the pool of passages is limited. The passages are supposed to reflect the mix of literary genres that the curriculum requires students to be familiar with – poems, informational pieces, fables, etc.
- A "test development team" is assembled – NYS Dept. of Education staff, teachers (retired and current), and CTB staff. This team reviews the passages to determine which ones are most appropriate for the test. This occurs even before the questions have been developed, because it's very expensive to develop the questions, and there's no point paying for development of questions for passages that are going to be rejected. Ms. Knight emphasized that "we considered the teachers the experts" at this stage – that the teachers on the team have the most weight in determining whether a particular passage is comprehensible, lends itself to good questions, etc. "If the teachers don't like them, we throw them out."
- For this year's ELA, I asked how many teachers were involved at this stage, and who they were. "Only a handful of teachers were involved, to be honest." She gave me the impression that most or all of them were retired, and that none were from New York City.
- I asked how it could happen that two of the three written passages on a single year's test could be fables about animals. She responded that although they try to use a mix of genres, they are limited by the terms of the contract with CTB; if they reject one passage, they can get stuck with an unrepresentative mix. "I know a lot of people don't like fables," she said.

- After passages are selected, CTB develops alternative sets of questions for each passage (another “very expensive” process), and brings them back for evaluation by another panel of teachers and Educ. Dept. staff. Again, she stressed that teachers carry a lot of weight at this stage. At this stage, the state is “stuck” with the passages, so the panel’s job is to choose among alternative sets of questions.
- In the case of the “Why the Rooster Crows at Dawn” questions on this year’s test, CTB first proposed a set of questions focused on the rooster, and how he changes in the story. The state’s team rejected those questions because, she said, they required students to analyze changes in the rooster’s thinking, not his outward behavior, which was too complex or ambiguous for fourth graders. The team preferred the alternative set of questions about Brownie the Cow, she said, because they were about the cow’s outward behavior, not her thinking.

[I argued otherwise on this point. Doesn’t the rooster change “outwardly”(e.g., it crows every morning), and aren’t the cow’s changes “inward” – related to her mood? Well... maybe. But the point of the questions, she said, is just to “get students writing.” You answer these questions like a lawyer, she said. You pick a side and make your case as best you can; the graders are supposed to evaluate “how you express yourself,” not the content of what you say.]

[I said that this private advice was helpful, but that it isn’t part of the standard test-prep or instructions to students, and that students who find a question illogical may not be able to confidently bullshit their way through an answer. “The smart kids and the analytical kids have problems with these questions,” she said. “They drive themselves crazy looking for the right answer.” The kids who “just start writing” do far better, she said.]

[But Ms. Knight did say that the “right” answer to the Brownie question is that the cow “starts out nice, and becomes mean.” I argued that Brownie and the other cows didn’t seem “mean” at all; they just played a gentle joke, and laughed to themselves. She said again that it’s about the writing, not the “right” answer.]

[Why did the questions have to be about one of the characters changing? I asked. Because, she said, “we have to ask how a character changes” – since character development is a major component of the curriculum. I suggested that fables are not the best vehicles for analyzing character development. She acknowledged that she had heard similar complaints repeatedly, but that the state is limited by the small pool of passages offered by CTB.]

[Ms. Knight also acknowledged problems with the essay questions comparing the two passages about the sun and the moon. “I’m not crazy about those,” she said.]

- If the team rejects one set of questions, it’s pretty much stuck with CTB’s alternatives, since, again, it’s so expensive and time-consuming to develop questions. She said (repeatedly): “We do the best we can under the circumstances” – the terms of the contract, deadlines, available personnel, etc.
- Somewhere along the way, the questions are field tested by students.
- Grading is done by local jurisdictions with guidance from the state and CTB. Ms. Knight reported that CTB staff from Indianapolis were brought in this year to “train the trainers,” who then train personnel (usually teachers, but sometimes “whoever they can get”) recruited by local school systems. The state provides graders with model essays and “consistency sets” (not sure what this is) to try to ensure uniform grading. Three graders review each test (not sure what happens if they disagree on writing sections). [From our later discussions with two teachers who worked as graders on this test, it appears that in practice there are sometimes only two graders reviewing each test. Each group has a leader who resolves scoring disagreements that come up between the two graders.] Ms. Knight acknowledges that there is no way to ensure complete uniformity, and that one group can end up grading similar-level essays much differently than another.
- In addition to the sample or model responses, graders are provided with written “rubrics” specific to each test question, giving guidance on how to evaluate responses.
- In evaluating the writing questions, Ms. Knight said the graders were instructed to “look at the writing holistically, look at all the writing together” – not just at one essay. [This was her response to more questioning about the effect of flawed questions on the test. I must say that I’m skeptical that graders would disregard “wrong” answers and focus exclusively on the quality of expression.]
- After the grading, then it gets really mysterious: “scaling” the test, or translating the “raw” scores (23 answers right) into a “scale” score (550). 100 or so teachers are involved, along with CTB and the state. This panel evaluates how/whether students were able to answer each question, and then gives each question relative weight (this one’s worth 15 points, that one’s worth 5). Ms. Knight gave me the impression there is lots of room for subjectivity here.
- The last step is setting the “cut scores”, deciding where to draw the lines between 1, 2, 3 and 4. Here again, a lot of mystery, variability and discretion.

“They do it differently every year,” she said, because of “different panels, different designs.”

- Finally, there’s an outside audit of the whole process. This year it’s being done by a firm called Pearson (?), but it’s not done yet.

Ms. Knight emphasized that Commissioner Mills and the state do not advocate the use of this test for anything other than its intended purpose. “A test is only valid for a specific purpose.” And, “we prefer multiple instruments to assess students.”

So what is the purpose of the test? “It’s really about the ones and twos,” Ms. Knight said – it’s designed to identify the children who are at risk of not meeting state standards. She seemed surprised (though I’m surprised that she wouldn’t know this) that NYC uses the test for a variety of other purposes, including middle-school admissions. She gave me the distinct impression that she thought this was a misuse of the ELA test.

She said that she was unaware of the decline in “4” scores in New York City, and that there was no effort by the state to evaluate this decline. She read me some of the statistics from the Power Point presentation given by Commissioner Mills when he released the aggregate results last week; these emphasized the number of kids “passing” – scoring 3 or 4 – without mentioning the percentage of kids scoring a 4.

I asked Ms. Knight what could account for the 60% decline in the percentage of NYC students scoring a 4. She said she thought it could have been the pressure. “I’ve never seen people so rattled as this year” – with reports of cheating, anxieties about teachers’ and principals’ jobs depending on scores, etc. When I asked her whether she thought an especially poorly designed test might contribute to the decline, she acknowledged that that was a possibility. She said the state wouldn’t automatically do a statistical analysis of the decline, but that it was within the power of Assistant Commissioner David Abrams to order one. It would be conducted by a psychometrician at the state.

Bottom line, she said: these are state tests, CTB constructs them, the feds (“who monitor everything carefully”) are “happy with them.” The whole process involves research and statistics, but in the end, it’s “an art form.”

Overall:

- Ms. Knight seemed aware that many people consider this year’s ELA test badly flawed. She didn’t defend the test much, but did defend the process – particularly the fact that *teachers* were involved in key roles at every step.
- Ms. Knight seemed to me to be disturbingly comfortable with essay questions that “drive smart kids crazy” because there’s no right answer.

- This year's writing portion of the test is, hopefully, about as bad as it gets. Yet it's hard to believe that this process used by the state could *ever* produce a test that gives a meaningful measure of children's writing, even in years when the state manages to avoid a Brownie-the-Cow-type fiasco.
- Above all, the New York City schools have no business using this test as an all-purpose evaluative tool. As flawed as these tests seem to be, and as misguided as the federal No Child Left Behind Law may be, no one is *making* Chancellor Klein use these tests in the way the city continues to use them. And no one would stand in the way if he decided to stop misusing them.

Joe Morris
Brooklyn, NY
October 4, 2006