Further Comment on "Lake Woebegone," Twenty Years Later

by Richard P. Phelps

Last summer, I searched out Dr. Cannell and asked if he would be willing to write an essay for our journal. He asserted that he had completely removed himself from the testing issue for twenty years and could not reliably write a research paper about the testing situation now. I countered that I had no intention of imposing that kind of burden on him. A short essay recounting his experiences of twenty years ago, when he wrote his two famous "Lake Wobegon" reports, was the sort of thing I had in mind.

The essay he sent to us was fine. With the exception of one short digression, it consisted entirely of his recollections, perfectly appropriate for posting as a commentary or op/ed piece.

Earnest, innocent manuscript put through hell

The editor at the time, however, insisted that it be sent out for review, as if anyone besides John J. Cannell would be qualified to review John J. Cannell's recollections. By the end of a several-month-long process, the paper had gone through several metamorphoses, grown several times in size, and became more of a research essay than a personal testimonial.

One reviewer reviewed versions one and two. Another reviewed only version two. Still another reviewed versions one and four. Around the time of version four, the latter reviewer was encouraged by the editor to contact Dr. Cannell and help him write the paper, presumably while still voting on the paper's acceptance.

The editorial handling of Dr. Cannell's paper was strange but, given the context of the unreasonable demands and constraints, Dr. Cannell's effort is admirable and substantial. And, even though our editorial review process was bungled, however, Dr. Cannell stands by the paper. I asked him. So, we have posted it.

Now, getting to testing

But, that's quite enough administrative background. Now, I address some of his weightier assertions, one by one.

Parallel tests as proof of cheating. The fact that trends in one test's scores do not parallel trends in another test's scores is not evidence of cheating. One is comparing apples and oranges--and two moving targets. At the very least, one should not even be thinking of making such a comparison unless one has first done a curriculum match study between the two tests. Saying that one can use trends in one test as evidence of cheating in another is about like saying that students who are training in, and improving their test scores in, oncology should be, measure for measure, improving their scores on podiatry tests, even though they aren't studying podiatry. (See, for example, Phelps, "The Source of Lake Wobegon", in this *Review*.)

Different standards as proof of cheating. Because West Virginia's test shows a higher percentage of "proficient" students than the NAEP does not prove cheating. It implies that West Virginia set its cut score for "proficient" relatively lower than did the NAEP but, again, these are apples and oranges. WV teachers are required to teach the West Virginia standards, not the NAEP standards.

Who’s to blame? As in his late-1980s Lake Wobegon reports, Dr. Cannell reserves a substantial part of the blame for the cheating he sees on testing companies. Perhaps it has something to do with my training in economics, but I tend to view businesses as morally inert– they allocate resources and supply demand. Testing companies did not inflate test scores then (or now), they sold products that could be used or misused by those who purchased them. Some, and only some, test users (i.e., state and local education officials) were responsible for the inflated test scores in the late 1980s.

There are only two types of tests–one good, one bad. There are three types of tests used in education: achievement, aptitude, and monitoring tests. Achievement tests are designed to measure how much you have learned about something in particular. Aptitude tests are designed to predict how well you might do in the future. Monitoring tests are designed to get a snapshot of an entire system's performance. The NAEP is a monitoring test. The ACT and SAT are "aptitude" tests whose sole reason for being is to predict future performance. (Yes, I know that they now call themselves “achievement” tests, but they focus on achievement that has a high predictive aspect.) They do this by measuring as wide an array of knowledge and skill as they can, with no particular attention paid to what the curriculum may be anywhere. The theory is that those who have the widest base of knowledge can most easily build new knowledge, say, in college.

It is not valid to compare the ACT and SAT to state achievement tests--they are different tests, designed for different purposes. By law, high stakes achievement tests (e.g., high school graduation tests) must be based on the standards and curriculum to which the students have been exposed. On these high stakes tests, often scores trend up over time, as they should. Students and teachers are motivated to work harder. Teachers learn to adapt their lessons over time to be more successful. Schools learn how to align their instruction over time to better match the standards on which the test is based. Meanwhile, scores on unrelated tests, not based on the same standards and curriculum may not rise in lockstep. And, why should they?

Broad-based tests. Dr. Cannell asserts that we should rely on national tests and also argues that our competitors overseas ruthlessly educate their children with broad-based tests. But, they don’t. The rest of the world does not use broad tests of achievement (except for, at best, a few percent of the total, and not then for high-stakes). The rest of the world uses standards-based, criterion-referenced, specific tests--mostly end-of-level and end-of-course tests--that are a 100% match to a jurisdiction-wide, uniform curriculum. "Broad tests of achievement" are almost

uniquely a North American development, derived from IQ and aptitude tests, and only imperfectly converted into achievement tests. Nationally norm-referenced, or "broad", tests of achievement, when taken “off the shelf,” are not usually well aligned with a particular state’s standards and so are unfair to use in high-stakes situations and, moreover, have been judged to be *illegal* to use in high-stakes situations.

The author recommends, using "broad tests of achievement", but that's what the Lake Wobegon tests were. The other, non-Lake Wobegon tests that Dr. Cannell encountered in the 1980s--the ones that were administered with high levels of security and item rotation--typically were standards-based and/or criterion-referenced tests (i.e., "narrow" and specific tests of achievement).

Tennessee math vs. American math. Like it or not, the U.S. Constitution says nothing about education. Therefore, the states are responsible for education. We can have national tests in the U.S., but they cannot be standards-based, criterion-referenced tests. They can, however, be Lake Wobegon tests. Like it or not, one can have West Virginia reading and West Virginia math, but there is no such thing as American reading or American math. States set standards, the U.S. does not. A state can have a uniform curriculum, the U.S. cannot. And, like it or not, state standards and state curricula vary substantially.

A few months ago, I tried my hand at writing mathematics test items. I gathered all the textbooks available to me and noticed (1) they teach a lot in the schools now that they did not teach when I was a kid (e.g., proofs in elementary school, discrete math (i.e., networking, graph theory, etc.), exploratory data analysis in elementary school, stats and probability in middle school) and (2) no two textbooks are alike. If one added up all the content in all the, say, 4th-grade textbooks, one would end up with 3 three years' worth of math instruction. No single school can teach all of it. Topics, and the sequencing of topics, vary from state to state and district to district. To expect students who have studied exploratory data analysis and graph theory to do just as well on a test that covers those two topics as they might on a different test that covers different topics (to which they have not been exposed) is unreasonable.

How much is enough? Dr. Cannell cites some state testing program characteristics as faults that I would characterize as virtues (e.g., a 50% item rotation rate is actually rather high, as some substantial number of items must be common from year to year so that scales can be equated from year to year). This makes we wonder if he considers any testing program to be a failure if it does not adhere to all of his criteria for good practice. That would make a testing program that adopted three-quarters of his criteria no better than one that adopted none. Dr. Cannell’s criteria for good practice are listed in his reports and on p.39 of my article in this same *Review*.

Nothing has changed. The author claims that "things have changed little since the 1980s," but he also admits that, by choice, he has not been following events. I believe that there is simply no comparison between, say, the California or Massachusetts testing programs of today and twenty years ago. Twenty years ago these states, essentially, did nothing, or pretty close to it.

Look at the hundreds of pages in their Web sites today, or the several dozen technical reports available, or the thousands of pages of data available, or the news coverage (mostly because now there are stakes to their tests).

This last point is the one that stands out in my mind the most, and most disappoints. Dr. Cannell argues that nothing has changed in twenty years. To my observation, everything has changed, and the situation is markedly improved. Moreover, things have changed and changed for the better in large part due to the heroic efforts of one person–Dr. John J. Cannell. I wish that he could appreciate that.