Comments on *"Lake Woebegone," Twenty Years Later* by J. J. Cannell, MD.

D. J. McRae, Ph.D.*
March 29, 2006

J. J. Cannell's article on the so-called "Lake Woebegone" effect for K-12 educational testing systems is mostly an historical account of technical issues and policy considerations that led in part to development of new types of test instruments for K-12 testing, i.e., standards-based tests. However, the article also comments on current testing practices, and charges that some of the technical issues and policy considerations that led to the Lake Woebegone effect are still in place. In particular, the Cannell article focuses on the California Standardized Testing and Reporting (STAR) program in recent years (pp 8-11 of the article).

In these comments, I'll focus on Cannell's comments on STAR. There perhaps is much of interest in the earlier historical treatment of the Lake Woebegone set of issues, but I'll resist muddying the water with those issues and focus more narrowly on the California STAR issues raised by Cannell.

In the article, Cannell argues that there are three testing system practices that affect the credibility of current STAR standards-based test results: (1) Many of the test questions are the same from year-to-year, allowing teachers to teach directly to test items rather than focus on teaching the content standards that underlie the test; (2) Test preparation materials are laced with test questions that unnecessarily narrow the curriculum; and (3) Teachers and school administrators administer the tests, thus allowing for outright cheating and/or undue influence on results.

Let me address these three practices in reverse order.

For practice # 3, Cannell suggests CA use outside or independent proctors to administer the STAR tests, like the standardized tests that Cannell cites early in the article [college entrance, medical school admissions, and medical licensure tests]. If money were not a consideration, I would find no fault with Cannell's argument. Certainly, using proctors who have a self-interest in the outcomes of the tests is a less than ideal practice, allowing for the possibility that the proctors outright cheat by helping students with the test (contrary to the test administration directions) or by changing answers after the fact or whatever. When instances of outright cheating do occur, they are widely publicized. However, the frequency of both intentional cheating as well as unintentional or inadvertent mistakes made during the test administration process has been quite low in K-12 testing programs. For the CA STAR program, with more than 200,000 test proctors, only 50 to 100 such instances are reported each year for investigation, only a handful are

confirmed as incidents that affect the validity of the test results, and in these cases the scores are nullified. The sanctions for teachers or school administrators involved in such irregularities are quite severe. Most folks agree that incidents of outright cheating by proctors of K-12 tests do not have such weight as to call for a major reform of this testing practice.

Nevertheless, it should be acknowledged that use of outside or independent proctors would be a testing practice that would improve the credibility of K-12 testing program results. But, such a practice would be extremely costly. The STAR program currently tests just fewer than 5 million students per year in grades 2-11 in CA at an operational cost of roughly $60 million per year, or $12/student. By way of contrast, the medical school admissions test that Cannell cites currently costs $210/student. Some of that additional cost may be traced to additional test forms being developed, and the psychometric overlap required to insure that the scores are comparable from test form to test form and from one test administration to another. But most of the increased cost is simply due to the administrative logistics needed to hire outside or independent proctors, to arrange for secure test sites, etc. The per student cost for the medical school admissions test is more than 17 times higher than the per student cost for STAR; if one requested a STAR budget from CA lawmakers to accommodate the testing practices found in the exam system cited by Cannell, the request would be a full ***billion*** (with a ***b***) dollars, rather than the current $60 million dollars. Such a request would be dead on arrival in the CA legislature.

It should be acknowledged there are less costly schemes to counteract potential cheating by proctors than the medical school admissions test scheme cited by Cannell. Indeed, one such practice is discussed as practice # 1 below. However, wholesale removal of teachers serving as test proctors for their own students would not be a measured public policy solution given the scope of this potential problem. The bottom line is that K-12 testing systems, for economic reasons, are designed with the assumption that teachers and school administrators will handle test proctoring duties honestly and ethically. And in the vast majority of cases, this is the case.

Practice # 2 cited by Cannell, the use or overuse of test preparation materials too closely matched to actual test questions, is a more serious issue (in my mind) than the issue of outside proctors. Actually, this is less a testing system issue than an instructional practice issue at its core. I am fond of saying that the best test preparation practice for a standards-based test is good solid standards-based instruction. I should follow that statement by saying that good solid standards-based instruction does not involve heavy use of test preparation materials. The appropriate use of test preparation materials should be simply to familiarize students with the formats likely to be encountered on a test.

However, it should be acknowledged that one of the unintended instructional practices that surfaces far too frequently in schools involved in high stakes testing programs is the overuse of test preparation materials too closely matched to actual test questions. In CA, there is statutory language prohibiting the use of instructional materials designed to raise

test scores without raising the underlying achievement measured by the test. However, it is one thing to put such words in legislative language, and it is another thing to enforce such language. There are no testing police out there to check on instructional practice, and hence there are no real sanctions for misuse of test preparation materials.

The fundamental assumption for the design of high stakes K-12 testing systems is that teachers and school administrators will make honest, ethical, and appropriate use of test preparation materials.. Unfortunately, too often the line between appropriate use and inappropriate abuse is obscured in practice. For highlighting this practice, I give Cannell at least half credit, though I do not agree with the rhetoric he uses to describe the overall effect of this practice nor do I agree that abuse of test preparation materials totally invalidates the information that comes from a K-12 system such as STAR. Rather, I would agree with the language of the Attorney General from Oklahoma who Cannell cites later in his article, that K-12 testing systems "lend themselves to being compromised" by such practices. For the longer term, widespread acceptance of the proposition that good solid standards-based instruction yields far better results on high stakes tests than weak test prep oriented instruction will discourage Practice # 2.

Practice # 1, the repeated use of test questions from year-to-year, is a testing system practice that has both pros and cons. First, it should be noted that kids do not see the same test questions year after year. Kids move on (for the most part) to new grade levels each year, and thus are exposed to an entirely new set of questions on the CA STAR statewide tests each year. Kids who are retained do see repeated test questions, but research has shown that test questions administered more than 6 months apart do not have a sufficient memory factor to invalidate test results. But, Cannell is not concerned about kids seeing repeated test questions. He is concerned that teachers have access to test questions and those teachers then use this knowledge to teach individual test items to the new crop of kids they have the next year.

In one sense, this is again an instructional issue rather than a testing issue. In this sense, it is at its core an issue of honesty and ethics. But the testing system also has some responsibility here. In designing a testing system, the strongest way to provide for gain score data (i.e., year-to-year comparisons) is to use repeated test items. Repeated use of good test items is also economical. But repeated use of test questions does open the door for potential item exposure that can compromise the test results. The classic way a testing system can combat repeated use of individual test questions is to design multiple alternate forms for tests, forms that measure the same thing but use differing test questions. This is a routine practice with high stakes higher education admissions tests; it is a practice sometimes used with K-12 tests, but not often enough. For the STAR standards-based tests, it has not been used to date. Cannell is accurate when he indicates that roughly 50 percent of the questions from a given form of a STAR standards-based test are repeated the following year; it would be a better practice to have multiple alternate forms for STAR standards-based tests, and to randomly assign such forms to schools. Such a practice would decrease the potential that any given teacher would be administering a significant number of repeated questions each year, and thus discourage

teachers using their access to last year's tests to "teach to test items" and thus compromise the credibility of STAR test results.

Use of multiple test forms is also a cost issue. The cost is primarily for additional test development, not additional test administration (though there would be some marginal increase in test administration costs). The additional cost for routine use of alternate forms may increase the cost of STAR by perhaps 10 or 20 percent [test development costs tend to be "fixed" costs that are independent of how many students are tested; with the large number of students tested under the CA STAR system, test development costs may be amortized over the large number of students and thus do not increase overall testing program costs the same way that operational costs (such as use of outside proctors) increase overall costs]. For Practice # 1, I would agree with Cannell that the testing system can and should do more to discourage the unintended practices that may come from repeated use of test questions from year-to-year.

I would have one additional note on a detail in Cannell's paper. On page 9, Cannell cites STAR scores from the *Stanford Achievement Test* (SAT) from 1998 to 2002 and compares them to scores from the *California Achievement Test* (CAT) from 2003, and concludes that the reading and language scores "plummeted" in 2003. The scores compared are apples and oranges since they are based on publisher norms samples from differing years and for different tests. The STAR test vendor supplied a conversion table in 2003 to convert the *SAT* scores from 1998 to 2002 to the *CAT* score scale. When the converted scores are analyzed, one finds that the CA reading and language scores from the national normed portion of STAR increased modestly from 2002 to 2003, following the trend previously identified from 1998 to 2002.

Allow me to complete these comments with several additional observations.

First, while I disagree or perhaps partially agree with Cannell on the specific testing system practices he identifies, I should comment that there are additional testing practices specific to the STAR standards-based tests that could be improved. In particular:

> I would note that the year-to-year comparisons of scores could be put on a more robust basis by paying greater attention to the assumptions required for the annual "equating" study that is conducted to put subsequent year test forms on the same scale of measurement. This is a technical issue, but it greatly affects the interpretation of test score gains from the STAR system.
>
> The STAR standards-based tests do not permit comparisons from grade to grade [they do not have the technical property called "vertical scaling"]. The STAR system would be on stronger technical grounds, particularly for use for accountability system calculations involving year to year gain scores, if it was designed to incorporate vertical scaling.

> The STAR system could have better reporting for parents and the public, particularly via the availability of "exemplars" to illustrate what performance levels such as Below Basic, or Basic, or Proficient, or Advanced mean.
> The STAR system could be coordinated with the high school exit testing system (CAHSEE) and the English Language Development testing system (CELDT) to eliminate redundant testing.
> The STAR standards-based system could avoid using near random scores to distinguish between Far Below Basic and Below Basic performance levels, thus permitting accountability index scores to increase without documented increases in achievement levels.

This is not the time or space to go into these improvements in greater detail, but it is appropriate to mention there are additional testing practices that need attention to improve the CA STAR system.

Second, I cannot help but make a few comments on the broader issues mentioned in the historical portion of Cannell's paper:

> On page 2, Cannell seems to make the assumption that public school tests should have the same properties as other standardized tests he is familiar with, to wit college entrance exams, medical school admissions tests, medical licensure tests. As an educational measurement specialist with more than 35 years experience, I have to comment there are many brands of standardized tests with each built to have specified properties unique to their intended applications. The most widely used K-12 tests during the 60's and 70's and 80's (when Cannell wrote his two "Lake Woebegone" reports) and at least half of the 90's were nationally normed commercial tests (NRTs); they were built to have certain properties that were appropriate for their intended use. When high stakes accountability use became prominent in the mid- to late-90's, the NRTs have been gradually replaced by standards-based tests custom designed to measure each state's unique academic content standards. The standards-based tests have their own set of specified properties that differ in many ways from the NRT properties. And both of these types of tests have properties that differ from the college entrance or medical school admissions tests that Cannell cites. Assuming that all standardized tests are built with the same set of properties in mind is not an accurate assumption.
> On pages 3-4, Cannell seems to equate blatant cheating by adults [such as erasing wrong answers and marking correct answers after students have completed their tests] with practices such as abuse of test preparation materials. The two sets of activities are not in the same domain. Blatant cheating is dishonest, unethical, and illegal; poor instructional practice needs to be corrected, but is not in the same category as blatant cheating.
> On pages 7-8, Cannell accuses test publishers of collaborating with others who provide test preparation materials. I'd like to say Cannell's accusation is

> balderdash, but my own view is that sometimes test publishers have become involved in the publishing of test preparation materials to the detriment of their efforts to publish good testing materials. There is a current trend for test publishers to become involved in what are called "formative" testing systems; these are instructional tests (not accountability or high stakes tests) that can be misused by turning out inappropriate test preparation materials. As a former test publisher, I have substantial concerns about this recent trend.

In his article, Cannell laments that his challenge to the K-12 testing establishment in the 1980's went for naught. I would disagree. In fact, in my opening sentence for this set of comments, I gave partial credit to Cannell for the development of new types of testing instruments for the K-12 testing system in the U.S. in the 1990's. Cannell uses flamboyant rhetoric, with words like "cheat" and "corrupted" that tend to obfuscate underlying legitimate testing system issues that he raises. Perhaps such rhetoric is needed to get attention for the issues he raises, particularly mainstream media attention. But, if one discards the rhetoric, Cannell raises issues that need to be addressed in the design of K-12 testing systems in the United States, and in this context he provides a valuable voice. I would not agree with Cannell's characterization that America has a "corrupt testing infrastructure." Rather, I would offer that America has an imperfect testing infrastructure, one that earnestly attempts to serve multiple masters with quality testing systems, one that has its warts and blemishes at any given moment in time, but one that also improves over time, as is the American way.

**Doug McRae is an Educational Measurement Specialist who resides in Monterey, California. He earned a Ph.D. in Psychometrics from the University of North Carolina at Chapel Hill, and has served in various capacities in the K-12 testing industry for more than 35 years. He served as Vice-President for Publishing at CTB McGraw-Hill in the early 1990's, overseeing development of standardized tests administered to 15 to 20 million K-12 students annually. He also served as Senior Advisor for the initial development of the Standardized Testing and Reporting Program (STAR) standards-based tests in California in the late 1990's.*

Citation: McRae, D.J. (2006) "Comments on 'Lake Woebegone' Twenty Years Later, by J.J. Cannell, M.D." *Nonpartisan Education Review / Essays, 2*(2). Retrieved [date] from http://www.npe.ednews.org/Review/Essays/v2n2.pdf