

Appendix C
The Rocky Score-Line of Lake Wobegonⁱ
Richard P. Phelps

Welcome to Lake Wobegon, where all the women are strong, all the men are good-looking,
 and all the children are above average.

– Garrison Keillor, *A Prairie Home Companion*

John Jacob Cannell's late 1980s reports caught all U.S. states asserting that their students' average scores on national norm-referenced tests were "above the national average." The phenomenon was dubbed the "Lake Wobegon Effect," in tribute to the mythical radio comedy community of Lake Wobegon, where "all the children are above average."

What caused the Lake Wobegon Effect? In his first (1987) report, Cannell identified several suspects—educator dishonesty (i.e., cheating) and conflict of interest, lax test security, inadequate or outdated norms, inappropriate populations tested (e.g., low-achieving students used as the norm group, or excluded from the operational test administration), and teaching the test.

In a table that summarizes the explanations given for spuriously high scores, one CRESST researcher provided a cross-tabulation of alleged causes with the names of researchers who had cited them (Shepard, 1990, 16). Conspicuous in their absence from the table, however, were Cannell's two primary suspects—educator dishonesty and lax test security.

This research framework presaged what was to come. The Lake Wobegon Effect continued to receive considerable attention and study from mainstream education researchers, especially those at CRESST, but Cannell's main points—that educator cheating was rampant and test security inadequate—were dismissed out of hand in the education research literature, and persistently ignored thereafter. In statistical jargon, this is called "Left-Out Variable Bias" (LOVB).

For almost two decades now, CRESST researchers have insisted that high-stakes, and not educator cheating or lax security, are responsible for "artificial" test score gains. They identify "teaching to the test" (i.e., test preparation or test coaching) as the direct mechanism that produces this "test score inflation" (Crocker, 2005).

The reasoning goes like this: under pressure to raise test scores, teachers "narrow the curriculum" to comprise only subject matter that will be covered on the test and "teach to the test." Moreover, teachers reduce the amount of time devoted to regular instruction and, instead, focus on test preparation that can be subject-matter free. This behavior, CRESST argues, leads to unintended consequences: test scores rise, but students learn less.

Research conducted on this hypothesis calls it into question, and in fact concludes that teachers who spent more than a brief amount of time focused on test preparation do their students more harm than good. Their students score lower on the tests than do other students whose teachers eschew any test preparation beyond simple format familiarization (and, instead, use the time for regular subject-matter instruction) (see, for example, Moore, 1991; Palmer, 2002; Crocker, 2005; Camara, 2008). Moreover, students who know the specific content of prep tests beforehand tend study less, learn less, and score lower on final exams than those who do not (see, for example, Tuckman, 1994; Tuckman & Trimble, 1997).

Finally, the empirical evidence cited by CRESST researchers to support their high-stakes-cause-test-score-inflation claim is less than abundant, consisting ofⁱⁱ

- the famous late-1980s “Lake Wobegon” reports of John Jacob Cannell (1987, 1989), as they interpret them.
- certain patterns in the pre- and post-test scores from the first decade or so of the Title I Evaluation and Reporting System (Linn, 2000, 5, 6).
- the “preliminary findings” from a un-replicable experiment they conducted themselves in the early 1990s in an unidentified school district, with two unidentified tests, one of which was “perceived to be high stakes” (Koretz, Linn, Dunbar, Shepard, 1991).

How good is this evidence?

Many educators and testing opponents consider the Cannell reports alone ample proof of the “score inflationary” effects of high-stakes testing and propose banning its use entirely, arguing that results from accountability tests cannot be trusted.

Indeed, Cannell’s data provide very convincing evidence of artificial test score inflation. But, with the exception of one Texas test, none of those that Cannell analyzed had stakes. Rather, all but one of his Lake Wobegon tests were used for system monitoring and diagnosis, and carried no consequences for students or teachers.

Cannell’s reports provide brief mentions of some state standards-based tests that had high stakes. Cannell contrasted their tight test security with the lax test security typical for the no-stakes NRTs whose scores he analyzed. But, he did not analyze the scores or trend in scores on the high-stakes standards-based tests. The Lake Wobegon tests—the tests whose scores inflated artificially over time—were no-stakes tests (Phelps, 2005b).

But, being mostly or entirely under the control of education administrators, the NRTs could be manipulated and their resulting scores published, making the administrators look good. Cannell’s data show that states that generally low-performing states were more prone to NRT score inflation, perhaps because administrators there felt embarrassed by their states showing on other measures and strove to compensate (Phelps, 2005b).

As the score-inflated tests themselves had no stakes, then, how could stakes have inflated their scores? Only if the stakes attached to other tests somehow affected the administrations of the NRTs. The states of Mississippi, North Carolina, and Arkansas exhibited strong score inflation with their NRTs in Cannell’s data, and all three states had other testing programs that had high stakes (with high levels of test security for those programs). But, Cannell’s own state of West Virginia also had terribly inflated NRT scores and no high-stakes testing program. The same was true for the neighboring state of Kentucky (Phelps, 2005b).

Nonetheless, I decided to look further into the CRESST hypothesis. I surmised that if high stakes cause test score inflation, one should find

- direct evidence that test coaching (i.e., teaching to the test), when isolated from other factors, increases test scores.
- grade levels closer to a high-stakes event (e.g., a high school graduation test) showing more test score inflation than grade levels further away.

For-profit test preparation and coaching services now comprise a multi-million dollar industry in the United States and their presence arouses equity concerns, as wealthier students are better able to afford them. Test prep firms claim that their services are worth the expense but, with the exception of the studies the firms themselves conduct, the research literature shows little benefit to paid coaching beyond that of some test format familiarization, which is just as easily accomplished by students on their own, for free.

Most research on the effect of test preparation has focused on university admissions tests, however. So, for direct evidence that test coaching increases tests scores the curious reader is invited to consult the thorough coverage of this topic, written by one of its foremost authorities, in Wayne Camara's chapter on admission testing (Camara, 2008).

Do Grade Levels Closer to a High-Stakes Event Show Greater Test Score Gains?

According to CRESST's Shepard:

Sixty-seven percent of... kindergarten teachers... reported implementing instructional practices in their classrooms that they considered to be antithetical to the learning needs of young children; they did this because of the demands of parents and the district and state accountability systems (Shepard, 1990, 21).

At the time Shepard wrote this, there existed a ninth-grade test in New Jersey with high-stakes. All other U.S. state high-stakes tests, however, were high school graduation examinations. The CRESST researcher was suggesting that pressure to succeed in high school graduation testing was distorting instruction at the kindergarten level (see also Shepard & Smith, 1988).

In education research jargon, when some aspect of a test given at one grade level has an effect on school, teacher, or student behavior in an earlier grade, there exists a "backwash" (or, washback) effect. Some testing researchers have attempted to learn whether or not a high-stakes testing program has backwash effects (many do), whether the effects are good or bad, and whether the effects are weak or strong (see, for example, Cheng & Watanabe, 2004). At least a few, however, have also tried to quantify those backwash effects.

The Cornell University labor economist John Bishop (1997) has found backwash effects from high stakes in most of his studies of testing programs. Typically, the high-stakes tests are given in some jurisdictions as requirements for graduation from upper secondary school (i.e., high school in the United States). Bishop then compares student performance on a no-stakes test given years earlier in these jurisdictions to student performance on the same no-stakes test given years earlier in jurisdictions without a high-stakes graduation examination. His consistent finding, controlling for other factors: students in jurisdictions with high-stakes graduation examinations—even students several years away from graduation—achieve more academically than students in jurisdictions without a high-stakes graduation exam.

So, Bishop's findings would seem to support Shepard's contention that the high stakes need merely be present somewhere in a school system for the entire system to be affected?

Not quite. First, Bishop identifies only positive backwash effects, whereas Shepard identifies only negative effects. Second, and more to the point, Bishop finds that the strength of the backwash effect varies, generally being stronger closer to the high-stakes event, and weaker further away from the high-stakes event. He calculated this empirically, too.

Using data from the Third International Mathematics and Science Study (TIMSS), which tested students at both 9- and 13-years old, he compared the difference in the strength of the backwash effect from high-stakes secondary school graduation exams between 13-year olds and 9-year olds. The backwash effect on 13-year olds appeared to be stronger in both reading and mathematics than it was on 9-year olds, much stronger in the case of mathematics. This suggests that backwash effects weaken with distance in grade levels from the high-stakes eventⁱⁱⁱ (Bishop, 1997, 10, 19).

This seems logical enough. Even if it were true that kindergarten teachers feel "high stakes pressure" to "teach the test" because the school district's high school administers a graduation

test, the pressure on the kindergarten teachers would likely be much less than that on high school, or even middle school, teachers.

Minimum-competency tests are high-stakes tests that require performance at or above a single threshold test score before certain educational attainment will be recognized. In a study of backwash effects of high school graduation exams on National Assessment of Educational Progress (NAEP) Reading scores, Linda Winfield, at the Educational Testing Service (ETS) wrote: “No advantages of MCT [minimum competency testing] programs were seen in grade 4, but they were in grades 8 and 11.” The presence-of-minimum-competency-test effect in grade 8 represented about an 8 point (.29 s.d. effect size) advantage for white students and a 10 point (.38 s.d. effect size) advantage for blacks in mean reading proficiency as compared to their respective counterparts in schools without MCTs. At grade 11 (after the point when most students have already taken their graduation exam), the effect represented a 2 point (.06 s.d. effect size) advantage for white students, a 7 point (.26 s.d. effect size) advantage for blacks, and a 6 point (.29 s.d. effect size) advantage for Hispanics (Winfield, 1990, 157).

The empirical evidence, then, disputes the assertion of CRESST’s Shepard that the pressure to succeed in high school graduation testing is translated into equivalent pressure in kindergarten. There might be some effect from high school graduation testing on the character of kindergarten in the same district. But, it is not likely equivalent to the effect that can be found at higher grade levels, nearer the high-stakes event.

Do Cannell’s data corroborate? Cannell (1989, 8, 31) himself noticed that test score inflation was worse in the elementary than in the secondary grades, suggesting that test score inflation declined in grade levels closer to the high-stakes event. I examined the norm-referenced test (NRT) score tables for each state in Cannell’s second report in order to determine the trend across the grade levels in the strength of test score inflation. That is, I looked to see if the amount by which the NRT scores were inflated was constant across grade levels, rose over the grade levels, or declined.

In over 20 states, the pattern was close to constant. In only two states did test scores rise with the grade levels, and they were both states without high-stakes testing. In 22 states, however, test scores declined as grade levels rose, and the majority of those states had high-stakes testing. (Cannell, 1989)

Why do Cannell’s data reveal exactly the opposite trend than the data from Bishop, Winfield, and Fredericksen? Likely, they do because the low-stakes test “control” in the two cases was administered very differently. Bishop, Winfield, and Fredericksen used the results from low-stakes tests that were administered both externally and to untraceable samples of students or classrooms. There was no possibility that the schools or school districts participating in these tests (e.g., the NAEP, the TIMSS) could or would want to manipulate the results.

Cannell’s Lake Wobegon tests were quite different. They were typically purchased by the school districts themselves and administered internally by the schools or school districts themselves. Moreover, as they were administered systemwide, there was every possibility that their results would be traceable to the schools and school districts participating. With the Lake Wobegon tests, the schools and school districts participating both could and would want to manipulate the results.

It would appear, then, that when tests are internally administered, their results can be manipulated. And, the farther removed these Lake Wobegon tests are (by grade level and, probably, by other measures) from the more high-profile and highly-scrutinized high-stakes tests, the more likely they are to be manipulated.

Conversely, it would appear that proximity to a high-stakes event (by grade level and, probably, by other measures) promotes genuine, non-artificial achievement gains.

The Elephants Not in the Room

Testimony that Cannell solicited from hundreds of educators across the country reinforced his wealth of empirical evidence in support of the notion that educator dishonesty and lax test security were constant companions of test score inflation, and the lower the stakes of a test, the more lax security tended to be (Cannell, 1989, chapter 3).

Since Cannell's reports provide no evidence that high stakes cause test score inflation, the empirical support for the CRESST hypothesis would seem to depend on their own preliminary study, which was conducted in an unnamed school district with unknown tests, one of which was allegedly perceived to be high stakes (Koretz, et al., 1991), and their interpretation of trends in Title I testing (Linn, 2000).

Seemingly Permanent Preliminary Findings

Researchers at the Center for Research on Education Standards and Student Testing (CRESST) have long advertised the results of a project they conducted in the early 1990s as proof that high stakes cause test score inflation (Koretz, et al., 1991).

For a study containing the foundational revelations of a widespread belief system, it is unusual in several respects:

- The study, apparently, never matured beyond the preliminary or initial findings stage or beyond implementation at just “one of [their] sites”, but many educators, nonetheless, appear to regard the study not only as proof of the high-stakes-cause-test-score-inflation hypothesis, but as all the proof that should be needed.
- It was neither peer-reviewed nor published in a scholarly journal. It can be found in the Education Resources in Education (ERIC) database in the form of a conference paper presentation.
- To this day, the identities of the particular school district where the study was conducted and the tests used in the study are kept secret (making it impossible for anyone to replicate the findings).
- As is typical for a conference paper presentation, which must be delivered in a brief period of time, much detail is left out, including rather important calculations, the definitions of certain terms, the exact meaning of several important references, some steps in their study procedures, and, most important, the specific content coverage of the tests and the schools' curricula.
- The stakes of the “high-stakes” test are never specified. Indeed, the key test may not have been high-stakes at all, as the authors introduce it thusly:

The district uses unmodified commercial achievement tests for its testing program, which is perceived as high-stakes (Koretz, et al., 1991, 4).

It is not explained how it came to be perceived that way, why it came to be perceived that way, nor who perceived it that way. Moreover, it is not explained if the third grade test featured in their study had high stakes itself, or if the high stakes were represented instead by, say, a high school graduation test, which made the entire “testing program” appear to have high stakes even though no stakes were attached to the third grade test.

- The study strongly suggests that curricula should be massively broad and the same in every school, but the study is conducted only in the primary grades.

In Koretz' own words, here is how the 1991 study was conducted:

Through the spring of 1986, [the district] used a test that I will call Test C. Since then, they have used another, called Test B, which was normed 7 years later than Test C. (4). For this analysis, we compared the district's own results—for Test C in 1986 and for Test B in 1987 through 1990—to our results for Test C. Our Test C results reflect 840 students in 36 schools. (6)

The results in mathematics show that scores do not generalize well from the district's test [i.e., Test B] to Test C, even though Test C was the district's own test only four years ago and is reasonably similar in format to Test B. (that is, both Test C and Test B are conventional, off-the-shelf multiple choice tests.) (6)

In other words, the CRESST researchers administered Test C, which had been used in the district until 1986 (and was in that year, presumably, perceived to have high stakes) to a sample of students in the district in 1990. They compared their sample of students' performance on this special, no-stakes test administration to the district's average results on the current high-stakes test, and they find differences in scores.

Why Should Different Tests Get the Same Result?

Why should it surprise anyone that students perform differently on two completely different, independently-developed norm-referenced tests (NRTs), and why should they care? Why should two different tests, developed by two completely different groups of people under entirely separate conditions, and using no common standard for content, be expected to produce nearly identical scores^{iv}?

Why should it surprise anyone that the primary school mathematics teachers in the unidentified large, urban school district taught different content and skills in 1990 than they did in 1986? Times change, curricula change, curricular requirements change, curricular sequencing changes, textbooks change, and, particularly in large, urban school districts, the teachers change, too.

Why should it surprise anyone that students perform better on a test that counts than they do on a test that does not?

I cannot answer these questions. But, the CRESST researchers, believing that the students should have scored the same on the different tests, saw a serious problem when they did not. From the abstract (Koretz, et al., 1991):

Detailed evidence is presented about the extent of generalization from high-stakes tests to other tests and about the instructional effects of high-stakes testing.... For mathematics, all comparisons, at district and student levels, support the primary hypothesis that performance on the conventional high-stakes test does not generalize well to other tests for which students have not been specifically prepared. Evidence in reading is less consistent, but suggests weaknesses in generalizing in some instances. Even the preliminary results presented in this paper provide a serious criticism of test-based accountability and raise concerns about the effects of high-stakes testing on instruction. Teachers in this district evidently focus on content specific to the test used for accountability rather than trying to improve achievement in the broader, more desirable sense.

This statement assumes (see the first sentence) that instructional behavior is the cause of the difference in scores, even though there were no controls in the study for other possible causes, such as variations in the stakes, variations in test security, variations in curricular alignment, and natural changes in curricular content over time.

Testing researchers who are not associated with CRESST caution against the simplistic assumptions that any test will generalize to any another with the same subject-area name and that one test can be used to benchmark trends in the scores of another (Cohen & Spillane, 1993, 53) Freeman et al. (1983) (Bhola, Impara, & Buckendahl, 2003, 28) (Archbald, 1994; Buckendahl, et al., 2000; Impara, et al., 2000; Plake, et al., 2000; Impara, 2001).

CRESST Response to Left-Out-Variable Bias

Koretz et al. (14, 15), do raise the topic of three other factors—specifically, variations in motivation, practice effects, and teaching to specific items (i.e. cheating). They admit that they “cannot disentangle these three factors” given their study design. Moreover, they admit that any influence the three factors would have on test scores would probably be in different directions.

Their attempt to account for these three factors they do identify was to administer a parallel form of Test B to a “randomly drawn” but unrepresentative sub sample of district third-graders. Scores from this no-stakes administration of the parallel Test B were reasonably consistent with the district scores from the regular administration of Test B. The CRESST researchers (14–18) cite this evidence as proof that motivation, practice effects, and possible teaching to specific items for the regular test administration have had no effect in this district.

This seems reassuring for their study, but also strange. In most experimental studies that isolate motivation from other factors, motivation exhibits a large effect on test scores (see, for example, Phelps, 2005a), but not in this study, apparently, as the sub sample of students score about the same on Test B (or, rather, somewhat higher on the parallel form), whether or not they took it under high- or no-stakes conditions. To my mind, the parallel-forms experiment only serves to resurface doubts about the stakes allegedly attached to the regular administration of Test B. If there genuinely were stakes attached to Test B at its regular administration, how can they have had no motivating effect? By contrast, if there were no stakes attached to Test B, the entire CRESST study was pointless.

Until the CRESST folk are willing to identify the tests they used in their little quasi-experiment, no one can compare the content of the two tests, and no one can replicate their study. No one’s privacy is at risk if CRESST identifies the two tests. So, the continued secrecy about the tests’ identities seems rather mysterious.

The Implications of “Teaching Away From the Test”

Another assumption in the statement from the study abstract seems to be that teachers are not supposed to teach subject matter content that matches their jurisdiction’s curricular standards (that would be “narrow”) but, rather, they are supposed to teach “more broadly” (i.e., subject matter that is outside their jurisdiction’s curricular standards). Leaving aside for the moment the issue of whether or not such behavior—deliberately teaching subject matter outside the jurisdiction’s curricular standards—would even be legal, where would it end?

Teachers are supposed to try to teach “broadly?” That’s not how it’s done in other countries. 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. And, each uniform curriculum can vary dramatically in content, and sequencing of content, from its counterparts in other townships, districts, states, provinces, cantons, communities, or nations. (See, for example, Robitaille, Schmidt, Raizen, McKnight, Britton, & Nicol, 1993; Howson, 1995; Robitaille, 1995; Schmidt, et al., 1996; Schmidt, McKnight, Valverde, Houang, Wiley, 1997.)

"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 (see Phelps, 2007, chapter 2).

Testing opponents are fond of arguing that scores from single test administrations should not be used for high-stakes decisions because the pool of knowledge is infinitely vast and any one standardized test can only sample a tiny fraction of the vast pool (see, for example, Heubert and Hauser, 1999, 3). The likelihood that one test developer's choice of curricular content will exactly equal another test developer's choice of curricular content is rather remote, short of some commonly-agreed upon mutual standard (i.e., something more specific and detailed than the National Council of Teachers of Mathematics Standards, which did not yet exist in 1990 anyway).

CRESST's Shepard, as a co-author of the 1991 Koretz et al. study, presumably would agree that average student scores from Test C and the five-year old Test B should be the same. But, curricula are constantly evolving, and five years is a long time span during which to expect that evolution to stop. In another context, Shepard (1990, 20) wrote:

At the median in reading, language, and mathematics [on an NRT], one additional item correct translates into a percentile gain of from 2 to 7 percentile points.

Shepard was trying to illustrate one of her claims about the alleged "teaching to the test" phenomenon. But, the point applies just as well to CRESST's insistence that scores on two different third-grade mathematics tests should correlate nearly perfectly. What if the first test assumed that third-graders will have been exposed to fractions by the time they take the test and the second test did not? What if the second test assumed the third-graders will have been exposed to basic geometric concepts, and the first test did not? What if the mathematics curricula everywhere had changed some over the five-year period 1986-1990? In any of these cases, there would be no reason to expect a very high correlation between the two tests, according to Shepard's own words displayed immediately above.

How Summer Vacation Deflates Test Scores

Another study sometimes cited as evidence of the high-stakes-cause-test-score-inflation hypothesis pertains to the pre-post testing requirement (or, Title I Evaluation and Reporting System (TIERS)) of the Title I Compensatory Education (i.e., anti-poverty) program from the late 1970s on. According to Linn (2000, 5):

Rather than administering tests once a year in selected grades, TIERS encouraged the administration of tests in both the fall and the spring for Title I students in order to evaluate the progress of students participating in the program.

Nationally aggregated results for Title I students in Grades 2 through 6 showed radically different patterns of gain for programs that reported results on different testing cycles (Linn, Dunbar, Harnisch, & Hastings, 1982). Programs using an annual testing cycle (i.e., fall-to-fall or spring-to-spring) to measure student progress in achievement showed much smaller gains on average than programs that used a fall-to-spring testing cycle.

Linn et al. (1982) reviewed a number of factors that together tended to inflate the estimates of gain in the fall-to-spring testing cycle results. These included such considerations as student selection, scale conversion errors, administration conditions, administration dates compared to norming dates, practice effects, and teaching to the test.

The last paragraph seems to imply that Linn et al. must have considered everything. They did not. For example, Title I testing of that era was administered without external quality control measures (see, for example, Sinclair & Gutman, 1992). Test security, just one of the influential factors not included in the Linn et al. list, was low or nonexistent.

Furthermore, Linn et al. (2000) did not consider the detrimental effect of summer vacation on student achievement gains. They assert that there are very different patterns of achievement gains between two groups: the first group comprises those school districts that administered their pre-post testing within the nine-month academic year (the nine-month cycle); and the second group comprises those school districts that administered their pre-post testing over a full calendar year's time (either fall-to-fall or spring-to-spring; the twelve-month cycle).

What is the most fundamental difference between the first and the second group? The pre-post testing for the first group involved no summer vacation or, rather, three months worth of forgetting; whereas the pre-post testing for the second group did include summer vacation, affording all the students involved three months to forget what they had learned the previous academic year.

True, Linn et al., considered several factors that could have influenced the outcome. However, they did not consider the single most obvious factor that could have influenced the outcome—the three-month summer layoff from study, and the deleterious effect that has on achievement gains.

Harris Cooper (1996) and others have reviewed the research literature on the effects of the summer layoff. According to Cooper:

The meta-analysis indicated that the summer loss equaled about one month on a grade-level equivalent scale, or one-tenth of a standard deviation relative to spring test scores. The effect of summer break was more detrimental for math than for reading and most detrimental for math computation and spelling (Cooper, 1996, abstract).

Given that the summer layoff more than compensates for the difference in scores between the first and second groups of Title I school districts, there seems little reason to pursue this line of inquiry any further. (It might be regarded as fairly obscure, anyway, that the difference in score gains between 12-month and 9-month pre-post testing cycles supports the notion that high stakes cause test score inflation.)

In summary, the high-stakes-cause-test-score-inflation hypothesis simply is not supported by empirical evidence.

Conclusion

The high-stakes-cause-test-score-inflation hypothesis would appear to be based on

- a misclassification of the tests in Cannell's reports (labeling the low-stakes tests as high-stakes);
- left-out variable bias;
- a cause-and-effect conclusion assumed by default from the variables remaining after most of the research literature on testing effects had been dismissed or ignored;
- a pinch of possible empirical support from a preliminary study conducted at an unknown location with unidentified tests, one of which was perceived to be high stakes; and
- semantic sleight-of-hand, surreptitiously substituting an overly broad and out-of-date definition for the term "high stakes".

The most certain cure for test score inflation is tight test security and ample item rotation, which are common with externally-administered, high-stakes testing. An agency external to the

local school district must be responsible for administering the tests under standardized, monitored, secure conditions, just the way it is done in hundreds of other countries. (See, for example, American Federation of Teachers 1995, Britton & Raizen 1996; Eckstein & Noah 1993; Phelps 1996, 2000, & 2001) If the tests have stakes, students, parents, teachers, and policy makers alike tend to take them seriously, and adequate resources are more likely to be invested toward ensuring test quality and security.

Any test can be made a Lake Wobegon test. All that is needed is an absence of test security and item rotation and the slightest of temptations for (some) educators to cheat. How a test is administered determines whether it becomes a Lake Wobegon test (i.e., one with artificial score gains over time). Ultimately, the other characteristics of the test, such as its name, purpose, content, or format, are irrelevant.

Two quite different test types prevent artificial test score gains (i.e., score inflation). One type has good security and ample item rotation, both of which are more common with high- than with low-stakes tests. The second type produces scores that are untraceable to schools or districts. Some system-monitoring and diagnostic tests bear this characteristic. Any test producing scores that are traceable to particular schools, districts, or states might also be used to make their administrators look good.

Experience shows that it does not take much incentive to induce at least some education administrators to cheat on standardized tests. But, cheating requires means, motive, and opportunity. When external agencies administer a test under tight security (and with ample item rotation), local school administrators are denied both means and opportunity to cheat. With tight security and item rotation, there can be no test score inflation.

The list that Cannell included in his 50-state survey of test security practices (1989, Appendix I) remains a useful reference. Jurisdictions wishing to avoid test score inflation should consider:

- enacting and enforcing formal, written, and detailed test security and test procedures policies;
- formally investigating all allegations of cheating;
- ensuring that educators cannot see test questions either before or after the actual test administration and enforce consequences for those who try;
- reducing as much as practicable the exclusion of students from test administrations (e.g., special education students);
- employing technologies that reduce cheating (e.g., optical scanning, computerized variance analysis);
- holding and sealing test booklets in a secure environment until test time;
- keeping test booklets away from the schools until test day;
- rotating items annually;
- prohibiting teachers from looking at the tests even during test administration;
- using outside test proctors; and
- spiraling different forms of the same test (i.e., having different students in the same room getting tests with different question ordering) to discourage student answer copying.

To Cannell's list from twenty years ago, one might add practices that consider the added advantages the Internet provides to those who cheat. Item rotation, for example, has become even more important given that any student can post (their recollection of) a test question on the Internet immediately after the conclusion of a test, thus aiding students taking the same test at a

later date or in a more westerly time zone the same day. Indeed, an entire company now exists that focuses solely on test security issues, specializing in Internet-related security problems.

References

- American Federation of Teachers. (1995). *Defining world class standards*. Washington, DC: Author.
- Archbald, D. (1994). *On the design and purposes of state curriculum guides: A comparison of mathematics and social studies guides from four states* (RR-029). Consortium for Policy Research in Education.
- Bhola, D. D., Impara, J. C., & Buckendahl, C. W. (2003, Fall). Aligning tests with States' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 21–29.
- Bishop, J. H. (1997). *Do curriculum-based external exit exam systems enhance student achievement?* (Paper 97-28). Ithaca, NY: Cornell University, Institute for Labor Relations, Center for Advanced Human Resource Studies.
- Britton, E. D., Hawkins, S., & Gandal, M. (1996). Comparing examinations systems. In E.D. Britton & S.A. Raizen (Eds.), *Examining the examinations: An international comparison of Science and Mathematics examinations for college-bound students* (pp. 201–218). Boston, MA: Kluwer Academic.
- Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M. (2000). *Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Camara, W. (2008). College admission testing: Myths and realities. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association.
- Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. (2nd Ed.), Daniels, WV: Friends for Education.
- Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.
- Cheng, L., & Watanabe, Y. (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.
- Cohen, D. K., & Spillane, J. P. (1993). Policy and practice: The relations between governance and instruction. In S. H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 35–95). San Francisco, CA: Jossey-Bass.
- Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996, Fall). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, 66(3), 227–268.
- Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 159–174). Mahwah, NJ: Lawrence Erlbaum.
- Eckstein, M. A., & Noah, H. J. (1993). *Secondary school examinations: International perspectives on policies and practice*. New Haven, CT: Yale University Press.
- Freeman, D., et al. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *Elementary School Journal*, 83(5), 501–514.
- Greene, J. P., Winters, M. A., & Forster, G. (2003). *Testing high-stakes tests: Can we believe the results of accountability tests?* (Report 33). Manhattan Institute, Center for Civic Innovation.

- Heubert, J. P., & Hauser, R. P. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Research Council.
- Howson, G. (1995). *Mathematics textbooks: A comparative study of grade 8 texts*. Vancouver, Canada: Pacific Educational Press.
- Impara, J. C. (2001, April). Alignment: One element of an assessment's instructional utility. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Impara, J. C., Plake, B.S., Buckendahl, C. W. (June, 2000). The comparability of norm-referenced achievement tests as they align to Nebraska's language arts content standards. Paper presented at the Large Scale Assessment Conference, Snowbird, UT.
- Keeghan, L. G. (2002). Comments for the panel, Accountability Systems that Work, at the Education Leaders Conference Annual Meeting, Denver, CO.
- Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND Education.
- Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April 5). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at R. L. Linn (Chair), *Effects of High-Stakes Educational Testing on Instruction and Achievement*, symposium presented at the annual meeting of the American Educational Research Association, Chicago.
- Linn, R. L. (2000, March). Assessments and accountability. *Educational Researcher*, 4–16.
- McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York, NY: Routledge.
- McNeil, L. M., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric*. Cambridge, MA: Harvard University, Civil Rights Project.
- Moore, W. P. (1991). *Relationships among teacher test performance pressures, perceived testing benefits, test preparation strategies, and student test performance*. PhD dissertation, University of Kansas, Lawrence.
- Palmer, J. S. (2002). *Performance incentives, teachers, and students: Estimating the effects of rewards policies on classroom practices and student performance*. PhD dissertation. Columbus, OH: The Ohio State University.
- Phelps, R. P. (1996, Fall). Are U.S. students the most heavily tested on Earth? *Educational Measurement: Issues and Practice*, 15(3), 19–27.
- Phelps, R. P. (2000). Trends in large-scale, external testing outside the United States. *Educational Measurement: Issues and Practice*, 19(1), 11–21.
- Phelps, R. P. (2001, August). Benchmarking to the world's best in Mathematics: Quality control in curriculum and instruction among the top performers in the TIMSS. *Evaluation Review*, 25(4), 391–439.
- Phelps, R. P. (2005a). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55–90). Mahwah, NJ: Lawrence Erlbaum.
- Phelps, R. P. (2005b). The source of Lake Wobegon. *Nonpartisan Education Review / Articles*, 1(2). <http://www.npe.ednews.org/Review/Articles/v1n2.htm>
- Phelps, R. P. (2007). *Standardized testing primer*. New York, NY: Peter Lang.
- Plake, B. S., Buckendahl, C. W., & Impara, J. C. (2000, June). A comparison of publishers' and teachers' perspectives on the alignment of norm-referenced tests to Nebraska's language

- arts content standards. Paper presented at the Large Scale Assessment Conference, Snowbird, Utah.
- Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science*. Vancouver, Canada: Pacific Educational Press.
- Robitaille, D. F. (1995). *National contexts for mathematics and science education*. Vancouver, Canada: Pacific Educational Press.
- Sandham, J. L. (1998, January 14). Ending SAT may hurt minorities, study shows. *Education Week*, 5.
- Schmidt, W. H., Jorde, D., Cogan, L. S., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G. A., McKnight, C., Prawat, R. S., Wiley, D. E., Raizen, S. A., Britton, E. D., & Wolfe, R. G. (1996). *Characterizing pedagogical flow: An investigation of mathematics and science teaching in six countries*. Boston, MA: Kluwer Academic.
- Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Boston, MA: Kluwer Academic.
- Shepard, L. A. (1990, Fall). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, 15–22.
- Shepard, L. A., & Smith, M. L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *The Elementary School Journal*, 89, 135–145.
- Sinclair, B., & Gutman, B. (1992). *A summary of state Chapter 1 participation and achievement information for 1989–90*. Washington, DC: U.S. Department of Education, Office of Policy and Planning.
- Smith, M. L. (1991a). *The role of testing in elementary schools* (CSE Technical Report 321). Los Angeles, CA: UCLA, Center for Research on Education Standards and Student Testing.
- Smith, M. L. (1991b, June). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5).
- Smith, M. L. (1991c, Fall). Meanings of test preparation. *American Educational Research Journal*, 28(3).
- Smith, M. L., & Rottenberg, C. (1991, Winter). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10–11.
- Thomson, J. A. (2000, October 26). Statement of Rand President and CEO James A. Thomson. (press release). Santa Monica, CA: Rand Corporation.
- Tuckman, B. W. (1994, April 4–8). *Comparing incentive motivation to metacognitive strategy in its effect on achievement*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, Available from ERIC (ED368790).
- Tuckman, B. W., & Trimble, S. (1997, August). *Using tests as a performance incentive to motivate eighth-graders to study*. Paper presented at the Annual Meeting of the American Psychological Association, Chicago, IL, Available from ERIC (ED418785).
- Winfield, L. F. (1990). School competency testing reforms and student achievement: Exploring a national perspective. *Education Evaluation and Policy Analysis*, 12(2), 157–173.

ⁱ This Appendix excerpts Phelps, R. P. (2005). The source of Lake Wobegon. *Nonpartisan Education Review / Articles*, 1(2). Available at

<http://www.npe.ednews.org/Review/Articles/v1n2.htm>

ⁱⁱ Some strikingly subjective (non-empirical) observational studies are sometimes cited as evidence, as well (see, for example, McNeil, 2000; McNeil & Valenzuela; Smith & Rottenberg; Smith 1991a–c).

ⁱⁱⁱ The difference in mathematics was statistically significant at the .01 level, whereas the difference in reading was not statistically significant.

^{iv} Regardless, the belief appears to be widespread. See, for example, Klein, Hamilton, McCaffrey & Stecher; Thomson; Keeghan; Greene, Winters & Forster.