THE SOURCE OF LAKE WOBEGON[1]

Richard P Phelps

[updated June 2010]

ABSTRACT

John J. Cannell's late 1980's "Lake Wobegon" reports suggested widespread deliberate educator manipulation of norm-referenced standardized test (NRT) administrations and results, resulting in artificial test score gains. The Cannell studies have been referenced in education research since, but as evidence that high stakes (and not cheating or lax security) cause test score inflation. This article examines that research and Cannell's data for evidence that high stakes cause test score inflation. No such evidence is found. Indeed, the evidence indicates that, if anything, the absence of high stakes is associated with artificial test score gains. The variable most highly correlated with test score inflation is general performance on achievement tests, with traditionally low-performing states exhibiting more test score inflation—on low-stakes norm-referenced tests—than traditionally high-performing states, regardless of whether or not a state also maintains a high-stakes testing program. The unsupported high-stakes-cause-test-score-inflation hypothesis seems to derive from the surreptitious substitution of an antiquated definition of the term "high stakes" and a few studies afflicted with left-out-variable bias. The source of test-score inflation is lax test security, regardless the stakes of the assessment.

Introduction

We know that tests that are used for accountability tend to be taught to in ways that produce inflated scores.
– D. Koretz, CRESST 1992, p.9

Corruption of indicators is a continuing problem where tests are used for accountability or other high-stakes purposes.
– R.L. Linn, CRESST 2000, p.5

The negative effects of high stakes testing on teaching and learning are well known.
Under intense political pressure, test scores are likely to go up without a corresponding improvement in student learning… all tests can be corrupted.

[1] The author acknowledges the generous assistance and advice of four anonymous, expert reviewers and that of the author of the Lake Wobegon reports, John J. Cannell. Of course, none of these several individuals is in any way responsible for any errors in this article.

– L.A. Shepard, CRESST 2000

> High stakes… lead teachers, school personnel, parents, and students to focus on just one thing: raising the test score by any means necessary. There is really no way that current tests can simultaneously be a legitimate indicator of learning and an object of concerted attention.
> – E.L. Baker, CRESST 2000, p.18

People cheat. Educators are people. Therefore, educators cheat. Not all educators, nor all people, but some.

This simple syllogism would seem incontrovertible. As is true for the population as a whole, some educators will risk cheating even in the face of measures meant to prevent or detect it. More will try to cheat in the absence of anti-cheating measures. As is also true for the population as a whole, some courageous and highly-principled souls will refuse to cheat even when many of their colleagues do.

Some education researchers, however, assert that deliberate educator cheating had nothing to do with the Lake Wobegon effect. Theirs are among the most widely cited and celebrated articles in the education policy research literature. Members of the federally-funded Center for Research on Education Standards and Student Testing (CRESST) have, for almost two decades, asserted that high-stakes cause "artificial" test score gains. They identify "teaching to the test" (i.e., test prep or test coaching) as the direct mechanism that produces this "test score inflation."

The High-Stakes-Cause-Test-Score-Inflation Hypothesis

The empirical evidence they cite to support their claim is less than abundant, however, largely consisting of,

- first, a quasi-experiment they conducted themselves fifteen years ago in an unidentified school district with unidentified tests (Koretz, Linn, Dunbar, Shepard 1991),
- second, 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, pp.5, 6), and
- third, the famous late-1980s "Lake Wobegon" reports of John Jacob Cannell (1987, 1989), as they interpret them.

Since the publication of Cannell's Lake Wobegon reports, it has, indeed, become "well known" that accountability tests produce score inflation. Well known or, at least, very widely believed. Many, and probably most, references to the Lake Wobegon reports in education research and policy circles since the late 1980s have identified high stakes, and only high stakes, as the cause of test score inflation (i.e., test score gains not related to achievement gains).

But, how good is the evidence?

In addition to studying the sources the CRESST researchers cite, I have analyzed Cannell's data in search of evidence. I surmised that if high stakes cause test score inflation, one should find the following:

- 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;
- direct evidence that test coaching (i.e., teaching to the test), when isolated from other factors, increases test scores; and
- an association between stakes in a state testing program and test score inflation.

One could call this the "weak" version of the high-stakes-cause-score-inflation hypothesis.

I further surmised that if high-stakes alone, and no other factor, cause artificial test score gains, one should find no positive correlation between test score gains and other factors, such as lax test security, educator cheating, student and teacher motivation, or tightening alignment between standards, curriculum, and test content.

One could call this the "strong" version of the high-stakes-cause-score-inflation hypothesis.

John Jacob Cannell and the "Lake Wobegon" Reports

> 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*

> It is clear that the standardized test results that were widely reported as part of accountability systems in the 1980s were giving an inflated impression of student achievement.
> – R.L. Linn, CRESST 2000, p.7

In 1987, a West Virginia physician, John Jacob Cannell, published the results of a study, *Nationally Normed Elementary Achievement Testing in America's Public Schools*. He had been surprised that West Virginia students kept scoring "above the national average" on a national norm-referenced standardized test (NRT), given the state's low relative standing on other measures of academic performance. He surveyed the situation in other states and with other NRTs and discovered that the students in every state were "above the national average," on elementary achievement tests, according to their norm-referenced test scores.

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." The Cannell report implied that half the school superintendents in the country were lying about their schools' academic achievement. It further implied that, with poorer results, the other half might lie, too.

School districts could purchase NRTs "off-the-shelf" from commercial test publishers and administer them on their own. With no "external" test administrators watching, school and district administrators were free to manipulate any and all aspects of the tests. They could look at the test items beforehand, and let their teachers look at them, too. They could give the students as much time to finish as they felt like giving them. They could keep using the same form of the test year after year. They could even score the tests themselves. The results from these internally-administered tests primed many a press release. (See Cannell 1989, Chapter 3)

Cannell followed up with a second report (1989), *How Public Educators Cheat on Standardized Achievement Tests*, in which he added similar state-by-state information for the secondary grades. He also provided detailed results of a survey of test security practices in the 50 states (pp.50–102), and printed some of the feedback he received from teachers in response to an advertisement his organization had placed in *Education Week* in spring 1989 (Chapter 3).

Institutional Responses to the Cannell Reports

> The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use…. The improper use of tests, however, can cause considerable harm….
> – AERA, APA, & NCME 1999, p.1

> The Lake Wobegon controversy led many of the testing corporations to be more timely in producing new norms tables to accompany their tests.
> – M. Chatterji 2003, p.25

The natural response to widespread cheating in most non-education fields would be to tighten security and to transfer the evaluative function to an external agency or agencies—agencies with no, or at least fewer, conflicts of interest. This is how testing with stakes has been organized in hundreds of other countries for decades.

Steps in this direction have been taken in the United States, too, since publication of Cannell's Reports. For example, it is now more common for state agencies, and less common for school districts, to administer tests with stakes. In most cases, this trend has paralleled both a tightening of test security and greater transparency in test development and administration.

There was a time long ago when education officials could administer a test statewide and then keep virtually all the results to themselves. In those days, those education officials with their fingers on the score reports could look at the summary results first, before deciding whether or not to make them public via a press release. Few reporters then even covered systemwide, and mostly diagnostic, testing much less knew when the results arrived at the state education department offices. But, again, this was long ago.

Legislative Responses

Between then and now, we have seen both California (in 1978) and New York State (in 1979) pass "truth in testing" laws that give individual students, or their parents, access to the corrected answers from standardized tests, not just their scores.[2] The laws also require test developers to submit technical reports, specifying how they determined their test's reliability and validity, and they require schools to explain the meaning of the test scores to individual students and their parents, while maintaining the privacy of all individual student test results.

Between then and now, we have seen the U.S. Congress pass the Family Education Rights and Privacy Act (FERPA), also called the Buckley Amendment (after the sponsor, Congressman

[2] The original New York State law, the *Educational Testing Act of 1979*, was updated in 1996 to apply to computer-administered as well as paper-and-pencil tests. The California law was based on the court case, *Diana v. California State Board of Education,* 1970.

James Buckley (NY)), which gives individual students and their parents similar rights of access to test information and assurances of privacy. Some federal legislation concerning those with disabilities has also enhanced individual students' and parents' rights *vis à vis* test information (e.g., the Rehabilitation Act of 1973).

Judicial Responses

Between then and now, the courts, both state and federal, have rendered verdicts that further enhance the public's right to access test-related information. *Debra P. v. Turlington* (1981) (Debra P. being a Florida student and Mr. Turlington being Florida's education superintendent at the time) is a case in point. A high school student who failed a nationally-norm-referenced high school graduation examination sued, employing the argument that it was not constitutional for the state to deny her a diploma based on her performance on a test that was not aligned to the curriculum to which she had been exposed. In other words, for students to have a fair chance at passing a test, they should be exposed to the domain of subject matter content that the test covers; in fairness, they should have some opportunity to learn in school what they must show they have learned on a graduation test. In one of the most influential legal cases in U.S. education history, the court sided with Debra P. against the Florida Education Department.

A more recent and even higher profile case (*GI Forum v. Texas Education Agency* (2000)), however, reaffirmed that students still must pass a state-mandated test to graduate, if state law stipulates that they must.

Response of the Professions

Cannell's public-spirited work, and the shock and embarrassment resulting from his findings within the psychometric world, likely gave a big push to reform as well. The industry bible, the *Standards for Educational and Psychological Testing*, mushroomed in size between its 1985 and 1999 editions, and now consists of 264 individual standards (i.e., rules, guidelines, or instructions) (American Educational Research Association 1999, pp. 4, 5):

> "The number of standards has increased from the 1985 *Standards* for a variety of reasons.... Standards dealing with important nontechnical issues, such as avoiding conflicts of interest and equitable treatment of all test takers, have been added… such topics have not been addressed in prior versions of the *Standards*."

The *Standards* now comprise 123 individual standards related to test construction, evaluation, and documentation, 48 individual standards on fairness issues, and 93 individual standards on the various kinds of testing applications (e.g., credentialing, diagnosis, and educational assessment). Close to a hundred member & research organizations, government agencies, and test development firms sponsor the development of the *Standards* and pledge to honor them.

Any more, to be legally defensible, the development, administration, and reporting of any high-stakes test must adhere to the *Standards* which, technically, are neither laws nor government regulations but are, nonetheless, regarded in law and practice as if they were. (Buckendahl & Hunt 2005)

Education Researchers' Response to the Cannell Reports

There are many reasons for the Lake Wobegon Effect,
most of which are less sinister than those emphasized by Cannell.
– R.L. Linn, CRESST 2000, p.7

The Cannell Reports attracted a flurry of research papers (and no group took to the task more vigorously than those at the Center for Research on Education Standards and Student Testing (CRESST)). Most researchers concurred that the Lake Wobegon Effect was real—across most states, many districts, and most grade levels, more aggregate average test scores were above average than would have been expected by chance—many more.

But, what caused the Lake Wobegon Effect? In his first (1987) report, Cannell named most of the prime 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," Shepard (1990, p.16) provided a cross-tabulation of alleged causes with the names of researchers who had cited them. Conspicuous in their absence from Shepard's table, however, were Cannell's two primary suspects—educator dishonesty and lax test security. This research framework presaged what was to come, at least from the CRESST researchers. 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, and persistently ignored thereafter.

Semantically Bound

The most pervasive source of high-stakes pressure identified by respondents was media coverage.
– L.A. Shepard, CRESST 1990, p.17

In his second (1989) report, Cannell briefly discussed the nature of stakes in testing. The definition of "high stakes" he employed, however, would be hardly recognizable today. According to Cannell (1989, p.9),

> "Professor Jim Popham at UCLA coined the term, 'high stakes' for tests that have consequences. When teachers feel judged by the results, when parents receive reports of their child's test scores, when tests are used to promote students, when test scores are widely reported in the newspapers, then the tests are 'high stakes.'"

Researchers at the Center for Research on Education Standards and Student Testing (CRESST) would use the same definition. For example, Shepard (1990, p.17) wrote:

> "Popham (1987) used the term high-stakes to refer to both tests with severe consequences for individual pupils, such as non-promotion, and those used to rank schools and districts in the media. The latter characterization clearly applies to 40 of the 50 states [in 1990]. Only four states conduct no state testing or aggregation of local district results; two states collect state data on a sampling basis in a way that does not put the spotlight on local districts. [Two

> more states] report state results collected from districts on a voluntary basis. Two additional states were rated as relatively low-stakes by their test coordinators; in these states, for example, test results are not typically page-one news, nor are district rank-orderings published."

Nowadays, the definition that Cannell and Shepard attributed to Popham is rather too broad to be useful, as it is difficult to imagine a systemwide test that would not fit within it. The summary results of any systemwide test must be made public. Thus, if media coverage is all that is necessary for a test to be classified as "high stakes," all systemwide tests are high stakes tests. If all tests are high stakes then, by definition, there are no low-stakes tests and the terms "low stakes" and "high stakes" make no useful distinctions.

This is a bit like calling all hours daytime. One could argue that there's some validity to doing so, as there is at all times some amount of light present, from the moon and the stars, for example, even if it is sometimes an infinitesimal amount (on cloudy, moonless nights, for example), or from fireflies, perhaps. But, the word "daytime" becomes much diminished in utility once its meaning encompasses its own opposite.

Similarly, one could easily make a valid argument that any test must have some stakes for someone; otherwise why would anyone make the effort to administer or take it? But, stakes vary, and calling any and all types of stakes, no matter how slight, "high" leaves one semantically constrained.

To my observation, most who join height adjectives to the word "stakes" in describing test impacts these days roughly follow this taxonomy:

> High Stakes – consequences that are defined in law or regulations result from exceeding, or not, one or more score thresholds. For a student, for example, the consequences could be completion of a level of education, or not, or promotion to the next grade level or not. For a teacher, the consequences could be job retention or not, or salary increase or bonus, or not.
>
> Medium Stakes – partial or conditional consequences that are defined in law or regulations result from exceeding, or not, one or more score thresholds. For a student, for example, the consequences could be an award, or not, admission to a selective, but non-required course of study, or not, or part of a "moderated" or "blended" score or grade, only the whole of which has high-stakes consequences.
>
> Low Stakes – the school system uses test scores in no manner that is consequential for students or for educators that is defined in law or regulations. Diagnostic tests, particularly when they are administered to anonymous samples of or individual students, are often considered low-stakes tests.

The definitions for "high-stakes test" and "low-stakes test" in the *Standards for Educational and Psychological Testing* (1999) are similar to mine above[3]:

---

[3] Note that the following CRESST researchers were involved in crafting the *Standards*: E.L. Baker, R.L. Linn, and L.A. Shepard. Indeed, Baker was co-chair of the joint AERA-APA-NCME committee that revised the *Standards* in the 1990s.

> "High-stakes test. A test used to provide results that have important, direct consequences for examinees, programs, or institutions involved in the testing." (p.176)
>
> "Low-stakes test. A test used to provide results that have only minor or indirect consequences for examinees, programs, or institutions involved in the testing." (p.178)

Note that, by either taxonomy, the fact that a school district superintendent or a school administrator might be motivated to artificially inflate test scores—to, for example, avoid embarrassment or pad a résumé—does not give a test high or medium stakes. By these taxonomies, avoiding discomfit is not considered to be a "stake" of the same magnitude as, say, a student being denied a diploma or a teacher losing a job. Administrator embarrassment is not a *direct* consequence of the testing nor, many would argue, is it an *important* consequence of the testing.

By either taxonomy, then, all but one of the tests analyzed by Cannell in his late 1980s-era Lake Wobegon reports were *low stakes* tests. With one exception (the Texas TEAMS), none of the Lake Wobegon tests was standards-based and none carried any direct or important state-imposed or state-authorized consequences for students, teachers, or schools.

Still, high stakes or no, some were motivated to tamper with the integrity of test administrations and to compromise test security. That is, some people cheated in administering the tests, and then misrepresented the results.

Wriggling Free of the Semantic Noose

> The phrase, *teaching the test*, is evocative but, in fact, has too many meanings to be directly useful.
> – L.A. Shepard, CRESST 1990, p.17.

> The curriculum will be degraded when tests are 'high stakes,' and when specific test content is known in advance.
> – J.J. Cannell 1989, p.26

Cannell reacted to the semantic constraint of Popham's overly broad definition of "high stakes" by coining yet another term—"legitimate high stakes"—which he contrasted with other high-stakes that, presumably, were not "legitimately" high. Cannell's "legitimate high stakes" tests are equivalent to what most today would identify as medium- or high-stakes tests (i.e., standards-based, accountability tests). Cannell's "not legitimately high stakes" tests—the nationally-normed achievement tests administered in the 1980s mostly for diagnostic reasons—would be classified as low-stakes tests in today's most common terminology. (See, for example, Cannell 1989, pp.20, 23)

But, as Cannell so effectively demonstrated, even those low-stakes test scores seemed to matter a great deal to someone. The people to whom the test scores mattered the most were district and school administrators who could publicly advertise the (artificial) test score gains as evidence of their own performance.

Then and now, however, researchers at the Center for Research on Education Standards and Student Testing (CRESST) neglected to make the "legitimate/non-legitimate," or any other, distinction between the infinitely broad Popham definition of "high stakes" and the far more narrow meaning of the term common today. Both then and now, they have left the definition

of "high stakes" flexible and, thus, open to easy misinterpretation. "High stakes" could mean pretty much anything one wanted it to mean, and serve any purpose.

Defining "Test Score Inflation"

> Cannell's reports …began to give public credence to the view that scores on high-stakes tests could be inflated.
> – D.M. Koretz, et al. CRESST 1991, p.2

Not only can the definition of the term "high stakes" be manipulated and confusing, so can the definition of "test score inflation." Generally, the term describes increases (usually over time) in test scores on achievement tests that do not represent genuine achievement gains but, rather, gains due to something not related to achievement (e.g., cheating, "teaching to the test" (i.e., test coaching)). To my knowledge, however, the term has never been given a measurable, quantitative definition.

For some of the analysis here, however, I needed a measurable definition and, so, I created one. Using Cannell's state-level data (Cannell 1989, Appendix I), I averaged the number of percentage-points above the 50th percentile across grades for each state, for which such data were available. In table 1 below, the average number of percentage points above the 50th percentile is shown for states with some high-stakes testing (6.1 percentage points) and for states with no high-stakes testing (12.1 percentage points).

Table 1.

State had high-stakes test?	Average number of percentage points above 50th percentile
Yes (N=13)	6.1
No (N=12)	12.2
25 states had insufficient data	

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

At first blush, it would appear that test score inflation is not higher in high-stakes testing states. Indeed, it appears to be lower.[4]

The comparison above, however, does not control for the fact that some states generally score above the 50th percentile on standardized achievement tests even when their test scores are not inflated. To adjust the percentage-point averages for the two groups of states—those with high stakes and those without—I used average state mathematics percentile scores from

[4] But, it is statistically significant only at the .10 level, in a t-test of means.

the 1990 or 1992 National Assessment of Educational Progress (NAEP) to compensate.[5] (NCES, p.725)

For example, in Cannell's second report (1989), the percentage-point average above the 50th percentile on norm-referenced tests (NRTs) is +20.3 (p.98). But, Wisconsin students tend to score above the national average on achievement tests no matter what the circumstances, so the +20.3 percentage points may not represent "inflation" but actual achievement that is higher than the national average. To adjust, I calculated the percentile-point difference between Wisconsin's average percentile score on the 1990 NAEP and the national average percentile score on the 1990 NAEP—+14 percentage points. Then, I subtracted the +14 from the +20.3 to arrive at an "adjusted" test score "inflation" number of +6.3.

I admit that this is a rough way of calculating a "test score inflation" indicator. Just one problem is the reduction in the number of data points. Between the presence (or not) of statewide NRT administration and the presence (or not) of NAEP scores from 1990 or 1992, half of the states in the country lack the necessary data to make the calculation. Nonetheless, as far as I know, this is the first attempt to apply any precision to the measurement of an "inflation" factor.

With the adjustment made (see table 2 below), at second blush, it would appear that states with high-stakes tests might have more "test score inflation" than states with no high-stakes tests, though the result is still not statistically significant.

Table 2.

State had high-stakes test?	Average number of percentage points above 50th percentile (adjusted)
Yes (N=13)	11.4
No (N=12)	8.2
25 states had insufficient data	

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

These data at least lean in the direction that the CRESST folk have indicated they should, but not yet very convincingly.[6]

[5] First, each state percentile average NAEP score was subtracted from the national percentile average NAEP score (for that year). Second, this difference was then subtracted from the state's number of percentage points above (or below) the 50th percentile on national norm-referenced tests, as documented in Cannell's second report.

[6] It is statistically significant only at the .10 level, in a t-test of means, the t-statistic being +1.27.

Testing the "Strong" Version of the High-Stakes-Cause-Score-Inflation Hypothesis

> Research has continually shown that increases in scores… reflect factors other than increased student achievement.
> Standards-based assessments do not have any better ability to correct this problem.
> – R.L. Linn, CRESST 1998, p.3

As mentioned earlier, the "strong" test of the high-stakes-[alone]-cause[s]-test-score-inflation hypothesis requires that we be unable to find a positive correlation between test score gains and any of the other suspected factors, such as lax test security and educator cheating.

Examining Cannell's data, I assembled four simple cross-tabulation tables. Two compare the presence of high-stakes in the states to, respectively, their item rotation practices and their level of test security as described by Cannell in his second report, The next two tables compare the average number of percentage points above the 50th percentile (adjusted for baseline performance with NAEP scores) on the "Lake Wobegon" tests—a rough measure of "test score inflation"—to their item rotation practices and their level of test security.

Item Rotation

Cannell noted in his first report that states that rotated items had no problem with test score inflation. (Cannell 1987, p.7) In his second report, he prominently mentions item rotation as one of the solutions to the problem of artificial test score gains.

According to Cannell, 20 states employed no item rotation and 16 of those twenty had no high-stakes testing. Twenty-one states rotated items and the majority, albeit slight, had high-stakes testing. (see table 3 below)

Table 3.

	Did state rotate test items?	
State had high-stakes test?	yes	no
Yes	11	4
No	10	16
9 states had insufficient data		

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

Contrasting the average "test score inflation," as calculated above (i.e., the average number of percentage points above the 50th percentile (adjusted by NAEP performance)), between item-rotating and non-item-rotating states, it would appear that states that rotated items had less test score inflation (see table 4 below).[7]

[7] But, a t-test comparing means shows no statistical significance, even at the 0.10 level.

Table 4.

	Did state rotate test items?	
	yes	no
Average number of percentage points above 50th percentile (adjusted)	9.3	10.0
29 state had insufficient data	N=12	N=9

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

Level of Test Security

Cannell administered a survey of test security practices and received replies from all but one state (Cannell 1989, Appendix I). As Cannell himself noted, the results require some digesting. For just one example, a state could choose to describe the test security practices for a test for which security was tight and not describe the test security practices for other tests, for which security was lax,… or vice versa. Most states at the time administered more than one testing program.

I classified a state's security practices as "lax" if they claimed to implement only one or two of the dozen or so practices about which Cannell inquired. I classified a state's security practices as "moderate" if they claimed to implement about half of Cannell's list. Finally, I classified a state's security practices as "tight" if they claimed to implement close to all of the practices on Cannell's list.

These three levels of test security are cross-tabulated with the presence (or not) of high-stakes testing in a state in table 5 below. Where there was lax test security, only four of 19 states had high-stakes testing. Where there was moderate test security, only four of 14 states had high-stakes testing. Where there was tight test security, however, eight of ten states had high-stakes testing.

Table 5.

	What was the quality of test security in the state?		
State had high-stakes test?	Lax	Moderate	Tight
Yes	4	4	8
No	15	10	2
7 states had insufficient data			

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

Contrasting the average "test score inflation," as calculated above (i.e., the average number of percentage points above the 50th percentile (adjusted by NAEP performance)), between lax, moderate, and tight test security states, it would appear that states with tighter test security

tended to have less test score inflation (see table 6 below).[8]

Table 6.

	What was the quality of test security in the state?		
	Lax	Moderate	Tight
Average number of percentage points above 50th percentile (adjusted)	10.6	9.7	8.9
27 states had insufficient data	N=12	N=5	N=6

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

At the very least, these four tables confound the issue. There emerges a rival hypothesis—Cannell's—that item rotation and tight test security prevent test score inflation. In the tables above, both item rotation and tight test security appear to be negatively correlated with test score inflation. Moreover, both appear to be positively correlated with the presence of high-stakes testing.

Testing the "Weak" Version of the High-Stakes-Cause-Score-Inflation Hypothesis.

> The implication appears clear: students… are prepared for the high-stakes testing in ways that boost scores on that specific test substantially more than actual achievement in the domains that the tests are intended to measure. Public reporting of these scores therefore creates an illusion of successful accountability and educational performance.
>
> –D.M. Koretz et al. CRESST 1991, pp.2, 3

As introduced earlier, the "weak" test of the high-stakes-cause-test-score-inflation hypothesis requires us to find: 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, direct evidence that test coaching (i.e., teaching to the test), when isolated from other factors, increases test scores, and an association between stakes in a state testing program and test score inflation.

I analyze Cannell's data to test the first two points. Cannell gathered basic information on norm-referenced test (NRT) scores by state for the school year 1987-88, including grades levels tested, numbers of students tested, and subject areas tested, and the percent of students and/or districts scoring at or above the 50th percentile. Where state-level information was unavailable, he attempted to sample large school districts in a state.

A page for one state—South Carolina—is reproduced from Cannell's second report and displayed later in this article.

[8] Comparing the mean for "lax" to that for "tight" produces a t-statistic not statistically significant, even at the 0.10 level.

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

> 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.
> – L.A. Shepard, CRESST 1990, p.21

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, this is called 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.

Bishop's studies. 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 (see quote above) 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 event.[9] (Bishop 1997, pp.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.

---

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

ETS studies. 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) found: "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 (.29 s.d. effect size) point advantage for white students and a 10 (.38 s.d. effect size) point advantage for blacks in mean reading proficiency as compared to their respective counterparts in schools without MCTs. At grade 11, the effect represented a 2 (.06 s.d. effect size) point advantage for white students, a 7 (.26 s.d. effect size) advantage for blacks, and a 6 (.29 s.d. effect size) advantage for Hispanics. (Winfield 1990, p.1) One should keep in mind that many states allowed to students to take their high-school graduation examination as early as eighth grade; in some states, the majority of students had already passed their graduation exam before they reached grade 12.

Norm Fredericksen, also at ETS, calculated NAEP score *gains* between 1978 and 1986 at three levels (for 9-, 13-, and 17-year olds). He found a significant effect for the youngest students—a 7.9 percentage-point gain [the NAEP scale ranges from 0 to 500 points]—for students in high-stakes testing states. He also found a 3.1 percentage-point gain for 13-year olds in high-stakes states in the same duration, which should be considered an additive effect [because, presumably, these students had already absorbed the earlier gains by the beginning of the time period]. An additional 0.6 percentage points were gained by 17-year olds over the time period. (Fredericksen 1994)

The empirical evidence, then, disputes Shepard's assertion that the pressure to succeed in high school graduation testing is translated into equivalent pressure in kindergarten in the same school district. (Shepard & Smith 1988) There might be some effect, whether good or bad, 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.

Cannell's studies. Do Cannell's data corroborate? Cannell (1989, pp.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. But, in only two states could one see test scores rising as grade levels rose, 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.[10] (see table 7 below)

[10] Ironically, the CRESST researchers themselves (Linn, Graue, & Sanders 1990, figure 3) offer evidence corroborating the pattern. Their bar chart (i.e., figure 3 in the 1990 article) clearly shows that, as grade levels rise, and get nearer the high-stakes event of graduation exams, scores on the NRTs fall. If they were correct that high stakes cause test score inflation, just the opposite should happen.

Table 7.			
	Trend in test scores from lower to higher grades.		
State had high-stakes test?	downward	level	upward
Yes	13	4	0
No	9	17	2
5 states have no data			
SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.			

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.

Is There Direct Evidence That Test Coaching, When Isolated from Other Factors, Increases Test Scores?

> Repeated practice or instruction geared to the format of the test rather than the content domain can increase scores without increasing achievement.
> - L.A. Shepard, CRESST 1990, p.19

If it is true that externally-administered, highly-secure, high-stakes tests can be "taught to," we should be able to find evidence of it in the experimental literature—in studies that test the coaching hypothesis directly. The research literature (discussed below) reveals a consistent result: test coaching does have a positive, but extremely small, effect.

Two separate aspects of test preparation. Essentially, there are two aspects to test preparation— (1) format familiarity and (2) remedial instruction or review in subject matter

mastery. Since commercial test prep courses (like those of Stanley Kaplan and the Princeton Review) are too short to make up for years of academic neglect and, thus, provide inadequate remedial help with subject matter mastery, what should one think of their ability to help students with format familiarity?

The most rigorous of the test coaching experiments in the research literature controlled the maximum number of other possible influential factors. Judging from their results, the only positive effect left from test prep courses seemed to be a familiarity with test item formats, such that coached examinees can process items on the operational test form more quickly and, thus, reach more test items. In other words, those who are already familiar with the test item structure and the wording of the test questions can move through a test more quickly than can those for whom all the material is fresh. This information, however, is available to anyone for free; one need not pay for a test prep course to gain this advantage. (Powers 1993, p.30)

Test preparation company claims. The Princeton Review's advertising claims, in particular, go far beyond familiarizing students with test format of the ACT or SAT, however. The Princeton Review argues that one can do well on multiple-choice standardized tests without even understanding the subject matter being tested. They claim that they increase students' test scores merely by helping them to understand how multiple-choice items are constructed. Are they correct?

The evidence they use to "prove" their case is in data of their own making. (See, for example, Smyth 1990) The Princeton Review, for example, gives some students practice SATs, scores them, then puts them through a course, after which they take a real SAT. They argue that the second SAT scores are hugely better. Even if one trusts that their data are accurate, however, it does not subtract out the effect of test familiarity. On average, students do better on the SAT just by taking it again. Indeed, simply retaking the SAT is a far less expensive way to familiarize oneself with the test.

According to Powers (1993, p.29):

> "When they have been asked to give their opinions, less than a majority of coached students have said they were satisfied with their score changes—for example, 24% of those polled by Snedecor (1989) and 43% of those surveyed by Whitla (1988)."

Moreover, the test preparation companies do not provide benefit-cost calculations in their benefit claims. Any test preparation course costs money, and takes time. That time spent in a test preparation course is an opportunity lost for studying on one's own that could be more focused, directed, and useful. (Powers 1993, p.29)

Results of studies on test preparation. For decades, independent scholars have studied the effect of test preparation courses like those offered by Stanley Kaplan and the Princeton Review. Becker's (1990) meta-analysis of such studies, for example, found only marginal effects for test coaching for the SAT. Becker analyzed study outcomes in terms of some 20 study characteristics having to do with both study design and content of coaching studied. Like previous analysts, she found that coaching effects were larger for the SAT-M (i.e., the mathematics section of the SAT) than for the SAT-V (the verbal section of the SAT). She did not find that duration of coaching was a strong predictor of the effects of coaching. Instead, she found that of all the coaching content variables she investigated, "item practice," (i.e., coaching in which participants were given practice on sample test items) was the strongest influence on

coaching outcomes). (Becker)

Overall, Becker concluded that among 21 published comparison studies, the effects of coaching were 0.09 standard deviations of the SAT-V and 0.16 on SAT-M. That is, just 9 points for the Verbal and 16 points for the Math, on their 500 point scales. That's virtually nothing, and far, far less than Stanley Kaplan and the Princeton Review claim.

Research completed in November 1998 by Donald Powers and Donald Rock update the earlier studies of Becker and others with new data about the minimal effects of coaching on the revised SAT, which was introduced in 1994.[11]

In surveying the research literature on test coaching, Powers noticed two compelling trends: first, the more rigorous the study methodology, the smaller the effect found from commercial test preparation courses (1993, p.26) and, second (1993, p.26):

> "…simply doubling the effort… does not double the effect. Diminishing returns set in rather quickly, and the time needed to achieve average score increases that are much larger than the relatively small increases observed in typical programs rapidly approaches that of full-time schooling (Messick & Jungeblut, 1981). Becker (1991) also documented the relationship between duration of coaching and effects on SAT courses, noting a weaker association after controlling for differences in the kind of coaching and the study design."

Most test coaching studies find only small correlations with test score changes. Testing

---

[11] As described by Wayne Camara (2001), Research Director of the College Board:
"Results from the various analyses conducted in the Powers and Rock study indicate the external coaching programs have a consistent but small effect on the SAT I, ranging in average effect from 21 to 34 points on the combined SAT I verbal and math scores. That is, the average effect of coaching is about 2 to 3 percent of the SAT I score scale of 400 to 1600 (the verbal and math scales each range from 200 to 800 points). Often raw score increases may be the easiest to understand. When examining the actual increases of both coached and uncoached students we find that:

- "Coached students had an average increase of 29 points on SAT verbal compared with an average increase of 21 points for uncoached students. Coached students had an average increase of 40 points on SAT math compared with 22 points for uncoached students. The best estimate of effect of coaching is 8 points on verbal scores and 18 points on math scores.
- "Coached students were slightly more likely to experience large score increases than uncoached students. Twelve and 16 percent of coached students had increases of 100 points or more on verbal and math scores, respectively, compared with 8 percent for uncoached students (on both math and verbal scores).
- "About one-third of all students actually had no gain or loss when retesting. On the verbal scale, 36 percent of coached students had a score decrease or no increase when retesting. On the math scale, 28 percent of coached students had a decrease or no increase, compared with 37 percent of uncoached students.
- "Students attending the two largest coaching firms, which offer the largest and most costly programs, do fare somewhat better than students attending other external coaching programs, but again, the effects of coaching are still relatively small. The typical gains for students attending these firms were 14 and 8 points on verbal scores and 11 and 34 points on math scores (with an average increase of 10 points on verbal, 22 points on math, and 43 points on combined verbal plus math for the two major test preparation firms).
- "There are no detectable differences in scores of coached students on the basis of gender and race/ethnicity, and whether initial scores were high or low.
- "The revised SAT I is no more coachable than the previous SAT.

"The estimated effects of coaching reported in this study (8 points on verbal and 18 points on math) are remarkably consistent with previous research published in peer reviewed scientific journals, all of which are at odds with the very large claims by several commercial coaching firms." (see also Briggs; DerSimonian and Laird; Kulik, Bangert-Drowns, and Kulik; Messick and Jungeblut, Zehr)

opponents typically dismiss these studies by ignoring them or, if they cannot ignore them, by attributing the results to researchers' alleged self-interest.[12]

After investigations and sustained pressure from better business groups, the Princeton Review in 2010 voluntary agreed to pull its advertising claiming score increases from taking its courses (National Advertising Division, 2010).

---

[12] Some testing opponents would have us believe that *all* the studies finding only weak effects from test coaching are conducted by testing organizations. That assertion is false. Nonetheless, *some* of them have been conducted by testing organizations.

Can one trust the results of studies sponsored by the College Board, or conducted by the ETS or the ACT? Certainly, these organizations have an interest in obtaining certain study results. ETS and ACT staffs develop tests for a living. If those tests can be gamed, then they do not necessarily measure the knowledge and skills that they purport to, and a high score can be obtained by anyone with the resources to pay to develop the gaming skills.

Moreover, ETS is very careful about what it prints in its reports. ETS vets its publications laboriously, often even deleting material it considers valid and reliable. A common reason for deletion, to my observation, is to avoid any offense of the many movers and shakers in education, on all sides of issues, with whom they seek to maintain good relations.

In this, they are not unlike many in the business of disseminating education research including, I would argue, most of the education press.

That being said, ETS behaves in a far more open-minded manner than many organizations in education. It has chosen more than half of its most prestigious William H. Angoff Memorial Lecture presenters, for example, from among the ranks of outspoken testing critics.

ETS' Policy Information Center routinely pumps out reports critical of the testing *status quo*, in which ETS plays a central part. Paul Barton's report, *Too Much Testing of the Wrong Kind and Too Little of the Right Kind* (1999b*)*, for example, lambastes certain types of testing that just happen to represent the largest proportion of ETS's revenues, while advocating types that represent only a negligible proportion of ETS's business. Other Barton and ETS Policy Information Center publications are equally critical of much of what ETS does. (see, for example, Barton 1999a)

The most compelling testimony in favor of the validity of the College Board, ETS, and ACT test coaching studies, however, is provided by the studies themselves. They tend to be high-quality research efforts that consider all the available evidence, both pro and con, and weigh it in the balance. Any study conducted by Donald Powers (at ETS), for example, provides a textbook example of how to do and present research well—carefully, thoroughly, and convincingly.

Whereas the ETS and the ACT clearly have incentives to justify their test development work, the College Board's self interest is not as clear. The College Board only sponsors tests; it neither develops nor administers them. Moreover, it comprises a consortium of hundreds of colleges and universities with incentives to sponsor tests only so long as they remain useful to them.

The only folk at College Board who possibly could have an incentive to continue using a useless test would be its small psychometric staff, arguably to protect their jobs. Given the current sizzling job market for psychometricians, however, it seems doubtful that they would risk sacrificing their impeccable professional reputations (by tainting or misrepresenting research results) in order to defend easily replaceable jobs.

Is There An Association Between Stakes In a Testing Program and Test Score Inflation?

> Both common sense and a great deal of hard evidence indicate that focused teaching to the test encouraged by accountability uses of results produces inflated notions of achievement.
> – R.L. Linn, CRESST 2000, p.7

In the earlier section "Defining 'Test Score Inflation'" I assembled the table below that contrasts the presence (or not) of high-stakes testing in a state and the amount of "test score inflation" on its nationally norm-referenced tests (NRTs). "Test score inflation" is manifest in this table as the average number of percentage points above the 50th percentile, adjusted by state NAEP scores.

Table 8.

State had high-stakes test?	Average number of percentage points above 50th percentile (adjusted)
Yes (N=13)	11.4
No (N=12)	8.2
25 states had insufficient data	

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, Appendix I.

It would appear that states with high-stakes tests might have more "test score inflation" than states with no high-stakes tests, though the difference is not strong.[13]

Considering General Achievement Levels

To be fair, however, another consideration must be taken into account. The decision to implement a high-stakes testing program in the 1980s was not taken randomly; the states that chose to were, on average, characteristically different from those that chose not to. One characteristic common to most high-stakes testing states was generally low academic achievement. States that ranked low on universal measures of achievement, such as the National Assessment of Educational Progress (NAEP), were more inclined to implement high-stakes testing than states that ranked high on measures of achievement. One could speculate that "low-performing" states felt the need to implement high-stakes testing as a means of inducing better performance, and "high-performing" states felt no such need.

Figure 1 below compares the amount of "test score inflation" in a state with the average state NAEP percentile score, from the 1990 or 1992 NAEP Mathematics test. States with high-stakes testing are indicated with squares; states without high-stakes testing are indicated with diamonds.

[13] It is statistically significant only at the .10 level, in a t-test of means, the t-statistic being +1.27.

Figure 1.

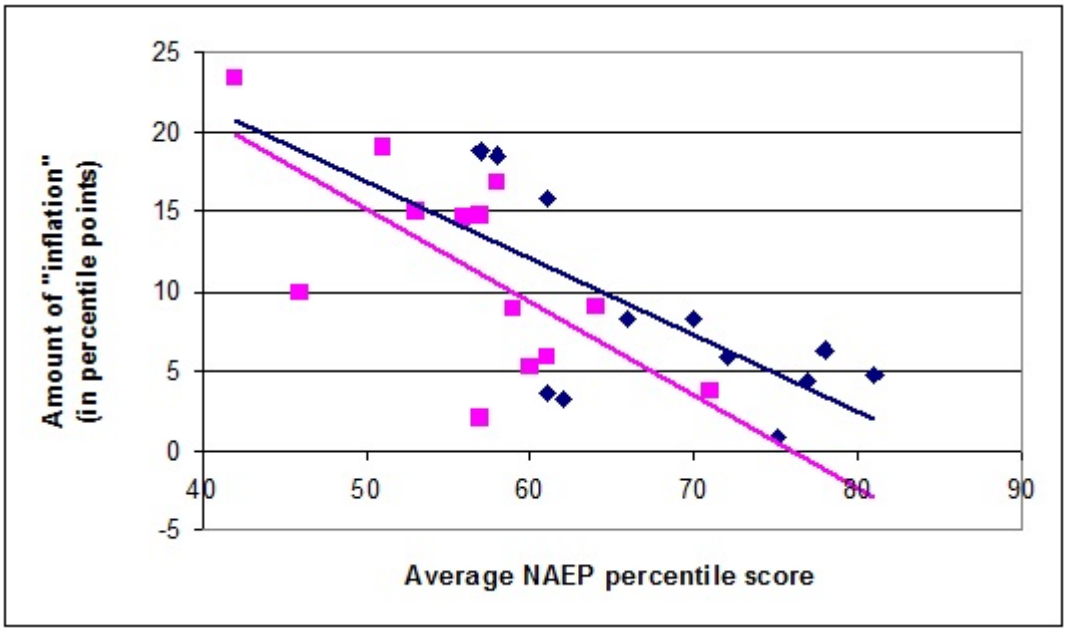

Figure 1 is revealing in several ways. First, a negative correlation between a state's general achievement level (as represented by average state NAEP percentile score) and its level of "test score inflation" is quite apparent. The Pearson product-moment correlation coefficient is –0.67, a fairly high correlation. It would appear that test score inflation is a function of a state's general achievement level—the lower a state's general achievement level, the higher the test score inflation is likely to be.

Second, figure 1 illustrates that generally low-achieving states are more likely to have high-stakes testing. One can see that the high-stakes states (the squares) tend toward the left side of the figure, whereas the other states (the diamonds) tend toward the right.

So, low-achieving states are more prone to implement high-stakes testing programs, and low-achieving states tend to exhibit more test score inflation (with their NRTs). If it were also true that high-stakes caused test score inflation, we might expect to see a higher fitted line through the high-stakes states (the squares in figure 1) than through the other states (the diamonds in figure 1).

We do not. The Pearson product-moment correlation coefficient for the high-stakes states is –0.68. The Pearson product-moment correlation coefficient for the low-stakes states is –0.65. Essentially, they are parallel, but the high-stakes line lies *below* the low-stakes line.

Multiple Regression

There are enough data to run a multiple regression of the test score inflation measure on the four factors considered thus far that are alleged to be correlated with test score inflation—item rotation, level of test security, presence of high stakes, and general state achievement level. No claims are made that this multiple regression is either elegant or precise. For one thing, only 20 of the 50 states have values for each of the four independent variables and the dependent variable as well. Nonetheless, as crude as it is, this analysis is far more sophisticated than any preceding it, to this author's knowledge.

Table 9 contains the multiple regression results.

Table 9. *Regression Statistics*

Multiple R	0.72
R Square	0.52
Adjusted R Square	0.39
Standard Error	4.88
Observations	20

ANOVA

	d.f.	*SS*	*MS*	*F*	*Sig. F*
Regression	4	385.47	96.37	4.05	0.0202
Residual	15	357.14	23.81		
Total	19	742.61			

	Coefficients	*Standard Error*	*t Statistic*	*P-value*
Intercept	45.70	10.20	4.48	0.0004
NAEP percentile score	-0.55	0.15	-3.72	0.0020
Item rotation (yes=1; no=0)	0.57	2.94	0.19	0.8501
Level of test security (tight=3; moderate=2; lax=1)	0.85	1.66	0.52	0.6140
High stakes (yes=1; no=1)	-6.47	3.51	-1.84	0.0853

Summarizing the results:

1) the data fit the function fairly well, with a multiple R statistic of 0.72;

2) the strongest predictor (significant at the 0.01 level) of test score inflation is NAEP percentile score (i.e., general achievement level), lending credence to a *new* theory that test score inflation is a deliberate, compensatory response on the part of education administrators to the publication of low achievement levels—the states with generally the lowest achievement, as shown on universal indicators such as the NAEP, exhibiting

the most of it; and

3) high stakes is the second strongest predictor, but it is statistically significant only at the 0.10 level and, more importantly, it has a *negative* sign, indicating that, if anything, the *absence of high stakes* is correlated with test score inflation.

It would seem that generally low-performing states tend to inflate their NRT scores, whether or not they have a high-stakes testing program. By all measures, Cannell's own state of West Virginia had terribly inflated NRT scores, but they had no high-stakes testing program. The same was true at the time for their neighboring state of Kentucky. Meanwhile, the states of Mississippi, North Carolina, and Arkansas also exhibited strong score inflation with their NRTs, but all three states had *other* testing programs that had high stakes and, also, high levels of test security for those programs.

Interpreting the results

This multiple regression offers a relatively decent test of the CRESST/high-stakes-cause-test-score-inflation hypothesis—the result being that the hypothesis must be rejected. We already know that the Lake Wobegon tests themselves were not high-stakes tests. Thus, the only way the CRESST hypothesis could be supported is if the mere "presence" of high-stakes testing in a state somehow led the officials responsible for the low stakes nationally norm-referenced (NRT) tests to inflate their test scores. The multiple regression results do not support such an allegation.

This multiple regression does not offer, however, a good test of Cannell's hypothesis—that the cause of test score inflation is lax test security and the educator cheating that takes advantage of it. First, we have no direct measure of educator cheating, so it can only be inferred. Second, the aforementioned problem with the returns from Cannell's 50-state survey of test security practices remains. That is, most states had multiple testing programs and, indeed, all but one of the states with a high-stakes testing program also administered a low-stakes testing program. Each respondent to the survey could choose the testing program for which the test security practices were described. The result is that some states may have conducted very lax security on their NRT programs, but very tight security for their high school graduation exams. A better test of Cannell's hypothesis would go through his data one more time attempting to verify which testing program's security practices were being described in the survey response, and then label only the test security practices for the NRTs (i.e., the Lake Wobegon tests).

Lynching the Most Disliked Suspect

> It is important to recognize the pervasive negative effects of accountability tests and the extent to which externally imposed testing programs prevent and drive out thoughtful classroom practices.… [*projecting image onto screen*] the image of Darth Vader and the Death Star seemed like an apt analogy.
> – L.A. Shepard, CRESST 2000

Thus far, we have uncovered strong evidence that test score inflation is (negatively) associated with states' general level of academic achievement and weaker evidence that test score inflation is (negatively) associated with the presence of high-stakes testing. Not only has the high-stakes-cause-test-score-inflation hypothesis not been supported by Cannell's data, the

converse is supported—it would appear that *low stakes* are associated with test score inflation. *Low reputation*, however, manifests the strongest correlation with test score inflation.

So, then, where is the evidence that high stakes cause test score inflation?

Some strikingly subjective observational studies are sometimes cited (see, for example, McNeil 2000, McNeil & Valenzuela 2000, Smith & Rottenberg 1991, Smith 1991a–c). But, the only *empirical* sources of evidence cited for the hypothesis that I know of are three: Cannell's "Lake Wobegon" reports from the late 1980s, patterns in Title I test scores during the 1970s and 1980s, and the "preliminary findings" of several CRESST researchers in a largely-secret quasi-experiment they conducted in the early 1990s with two unidentified tests, one of which was "perceived to be high stakes." (Koretz, et al. 1991)

Cannell's reports, however, provided statistics only for state- or district-wide nationally norm-referenced tests (NRTs). At the state level at least, the use of national NRTs for accountability purposes had died out by the mid-1980s, largely as a result of court edicts, such as that delivered in *Debra P. vs. Turlington.* The courts declared it to be unfair, and henceforth illegal, to deny a student graduation based on a score from a test that was not aligned with the course of study offered by the student's schools. From that point on, high-stakes tests were required to be aligned to a state's curricular standards, so that students had a fair chance to prepare themselves in the content domain of the test.[14]

Cannell's data provide very convincing evidence of artificial test score inflation. But, with the exception of one test in Texas—the TEAMS, which had been equated to the Metropolitan Achievement Test, an NRT—there were no accountability tests in Cannell's collection of tests nor were those tests "part of accountability systems." He does mention the existence of accountability tests in his text, often contrasting their tight test security with the lax test security typical for the NRTs, but he provides no data for them. Accountability tests are not part of his Lake Wobegon Effect.

[14] A more recent and even higher profile case (*GI Forum v. Texas Education Agency* (2000)), however, reaffirmed that students still must pass a state-mandated test to graduate, if state law stipulates that they must.

In Exhibit 1 below is an example of how Cannell (1989) presented his NRT information alongside that for accountability tests. For South Carolina (p.89), Cannell presents this table of results from statewide testing with the Comprehensive Test of Basic Skills (CTBS):

Exhibit 1.

SOUTH CAROLINA March 1989 Comprehensive Test of Basic Skills, Form U 1981 National Norms

Grade	Number tested	Reading	Language	Math	Total battery	% students > OR = 50	% districts > OR = 50
4	46,706	57.1	64.4	69.6	62.0	64.3%	81/92(88%)
5	45,047	51.6	59.3	66.9	55.2	55.5%	51/92(55%)
7	44,589	52.8	63.0	66.6	58.4	60.4%	71/92(77%)
9	47,676	49.1	58.3	60.2	54.2	53.8%	50/93(54%)
11	36,566	45.4	64.2	61.3	56.1	54.8%	48/93(52%)

Reporting method: median individual national percentiles
Source: *South Carolina Statewide Testing Program, 1989 Summary Report*.

TEST SECURITY IN SOUTH CAROLINA

South Carolina also administers a graduation exam and a criterion referenced test, both of which have significant security measures. Teachers are not allowed to look at either of these two test booklets, teachers may not obtain booklets before the day of testing, the graduation test booklets are sealed, testing is routinely monitored by state officials, special education students are generally included in all tests used in South Carolina unless their IEP recommends against testing, outside test proctors administer the graduation exam, and most test questions are rotated every year on the criterion referenced test.

Unlike their other two tests, teachers are allowed to look at CTBS test booklets, teachers may obtain CTBS test booklets before the day of testing, the booklets are not sealed, fall testing is not required, and CTBS testing is not routinely monitored by state officials. Outside test proctors are not routinely used to administer the CTBS, test questions have not been rotated every year, and CTBS answer sheets have not been routinely scanned for suspicious erasures or routinely analyzed for cluster variance. There are no state regulations that govern test security and test administration for norm-referenced testing done independently in the local school districts.

SOURCE: J.J. Cannell, *How Public Educators Cheat on Standardized Achievement Tests*, p.89.

The first paragraph in the test security section on South Carolina's page describes tight security for state-developed, standards-based high-stakes tests. There simply is no discussion of, nor evidence for, test score inflation for these accountability tests. The second paragraph describes the test with the inflated scores that are listed in the table at the top of the page. That test—the nationally norm-referenced CTBS—was administered without stakes (by today's definition of stakes) and, likewise, with lax test security. It—the low-stakes test—is the one that betrays evidence of test score inflation.

The rest of the state pages in Cannell's second report tell a similar story. The high-stakes tests were administered under tight security and there was no mention of test score inflation

in their regard. The low-stakes tests were sometimes, but not usually, administered under tight security and, when security was lax, test score inflation was usually present.

The Elephants in the Room

I believe in looking reality straight in the eye and denying it.
– G. Keillor, *A Prairie Home Companion*

Cannell's data do not show that accountability tests cause, or are even correlated with, test score inflation. Cannell pins the blame for test score inflation, first and foremost, on two culprits: educator dishonesty and lax test security.

The researchers at the Center for Research on Education Standards and Student Testing (CRESST), however, give little to no consideration in their studies to any of the primary suspects for test score gains—educator dishonesty and lax test security (usually when the stakes are low), curricular alignment and motivation (usually when the stakes are high), and generally low achievement levels, regardless the stakes. CRESST studies do not find that these factors lead to test score gains, because they do not consider these factors in their studies in the first place.

In statistical jargon, this is called "Left-Out Variable Bias" or, more affectionately, LOVB.

Testimony that Cannell solicited from hundreds of educators across the country reinforces 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 that lax test security is more common with low-stakes tests. (Cannell 1989, chapt.3)

As for high-stakes tests, there exist dozens of studies providing experimental and other empirical support for the notion that tightening the standards-curriculum-test alignment is associated with test score gains over time. Likewise, there exist hundreds of studies providing experimental and other empirical support for the notion that high-stakes-induced motivation is associated with test score gains over time. (see, for example, Phelps 2005, Appendix B)

CRESST researchers, to my knowledge, have done nothing to make their clients (the U.S. taxpayers) aware of these other research studies, with conclusions that contradict theirs. Even better, they sometimes *declare* that the hundreds of other studies do not exist. According to CRESST researcher D.M. Koretz (1996):

> "Despite the long history of assessment-based accountability, hard evidence about its effects is surprisingly sparse, and the little evidence that is available is not encouraging."

Likewise, a panel hired by the National Research Council (where CRESST researchers serve regularly as panel members) over a decade ago (Hartigan & Wigdor 1989), declared there to be no evidence of any benefit from the use of employment testing. This, despite the fact that over a thousand controlled experiments had been conducted finding those benefits to be pronounced and persistent. (Phelps 1999)

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

> We expected that the rosy picture painted by results on high-stakes tests would be to a substantial degree illusory and misleading.
> – D.M. Koretz, et al. CRESST 1991, p.1

> Even the preliminary results we are presenting today provide a very serious criticism of test-based accountability…. Few citizens or policy makers, I suspect, are particularly interested in performance, say, on "mathematics as tested by Test B but not Test C."
> They are presumably much more interested in performance in mathematics, rather broadly defined.
> – D.M. Koretz, et al. CRESST 1991, p.20

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 (not that peer reviewing means very much in education research) 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, some detail is absent, including rather important calculations, definitions of certain terms, meanings 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 1991, p.4) It is not explained how it came to be perceived that way, why it was perceived that way, nor who perceived it that way. Moreover, it is not explained if the third grade test itself had high stakes, or if the high stakes were represented instead by, say, a high school graduation test, which gave the entire "testing program" an appearance of 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.[15]

Study Design

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

> "The district uses unmodified commercial achievement tests for its testing program, which is perceived as high-stakes. Through the spring of 1986, they 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. (p.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. (p.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.)" (p.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 compare 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.[16]

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?

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

---

[15] Curricula can differ across schools for several reasons, including: differences in standards, differences in alignment to the standards, differences in the degree to which the standards are taken seriously, and differences in the sequencing in which topics are covered. Different schools can adhere equally well to content standards while sequencing the topics in entirely different orders, based on different preferences, different textbooks, and so on.

Schools are likely to modify their curricula, and their sequencing, to align them with a high-stakes standards-based test. Otherwise, their students will face curricular content on the test that they have not had an opportunity to learn, which would be unfair to the students (and possibly illegal for the agency administering the test). Conversely, schools are *unlikely* to modify their curricula, and their sequencing, to align them with a no-stakes NRT, particularly if they also administer a high-stakes standards-based test that assumes different content and different sequencing.

[16] The percentile ranks are listed as 42, 67, and 48 for, respectively, for the reading, mathematics, and vocabulary sections of Test B, and as 38, 51, and 35 for the same sections of Test C. The grade-equivalent scores are listed as 3.4, 4.5, and 3.6 for, respectively, the reading, mathematics, and vocabulary sections of Test B, and as 3.4, 3.8, and 3.4 for the same sections of Test C.

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.

CRESST Response to LOVB

Koretz et al., 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. (p.14) Moreover, they admit that any influence the three factors would have on test scores would probably be in different directions. (p.14)

Their solution to the 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. (p.15) 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 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. (pp.14-18)

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 2005), 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?

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, p.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 *Principles and Standards* (1991), which did not yet exist in 1990 anyway).

Teachers are supposed to try to teach the entirety of the possible curriculum? Third grade mathematics teachers, for example, are supposed to teach not only the topics required by their own jurisdiction's legal content standards, but those covered in any other jurisdiction, from Papua New Guinea to Tristan de Cunha? Any subject matter that is taught in third grade anywhere, or that has ever been taught in third grade anywhere, must be considered part of the possible curriculum, and must be taught? It could take several years to teach that much content.

L.A. 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, p.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 assumes that third-graders will have been exposed to fractions by the time they take the test and the second test does not? What if the second test assumes the third-graders will have been exposed to basic geometric concepts, and the first test does not? What if the mathematics curricula everywhere has 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.

Who Speaks for "The Public"?

In a quote at the outset of this section of the article, D.M. Koretz asserts that the public is not interested in students' performing well on a particular mathematics test but, rather, in all of mathematics. (Koretz, et al. 1991, p.20) I doubt that he's correct. Most everyone knows that the quantity of subject matter is boundless. No one can learn all the mathematics there is to learn, or even what is considered by various parties throughout the globe to represent *third-grade* level mathematics. Likewise, no one can learn all the mathematics that is covered in all the various third-grade mathematics textbooks, standards documents, curriculum guides, and so on.

More likely, what the public wants their third graders to learn is *some* coherent and integrated mathematics curriculum. I would wager that most Americans would not be picky about which of the many possible mathematics curricula their third-graders had learned, if only they could feel assured that their third-graders had learned one of them.

In their chapter of the book, *Designing Coherent Education Policy* (1993, p.53), David Cohen and James Spillane argue that:

> "Standardized tests often have been seen as interchangeable, but one of the few careful studies of topical agreement among tests raised doubts about that view. Focusing on several leading fourth grade mathematics tests, the authors observed that 'our findings challenge . . . th[e] assumption . . . that standardized achievement tests may be used interchangeably' (Freeman and others, 1983). The authors maintain that these tests are topically inconsistent and thus differentially sensitive to content coverage."

More recently, Bhola, Impara, and Buckendahl (2003) studied the curricular alignment of five different widely-available national norm-referenced tests for grades four and eight, and for high school, to Nebraska's state reading/language arts standards for grades four and eight, and for high school (p.28).

> "It was concluded that there are variable levels of alignment both across grades and across tests. No single test battery demonstrated a clear superiority in matching Nebraska's reading/language arts standards across all standards and grade levels. No test battery provided a comprehensive assessment of all of Nebraska's reading/language arts content standards. The use of any of these tests to satisfy NCLB requirements would require using additional assessment instruments to ensure that all content standards at any particular grade level are appropriately assessed.…
>
> "Our findings are consistent with those of La Marca et al. (2000) who summarize the results of five alignment studies that used different models to determine degree of alignment. In general, all these alignment studies found that alignments between assessments and content standards tended to be poor."

"Generalizability" Across Different Content Standards?

The CRESST folk (Koretz, et al. 1991), as well as Freeman, et al. (cited by Cohen and Spillane

above) and Bhola, Impara, and Buckendahl (2003), used "off-the-shelf" norm-referenced tests (NRTs) as points of comparison. But, what would become of CRESST's argument about "generalizability" if the tests in question had been developed from scratch as standards-based tests (i.e., with different standards reference documents, different test framework writers and review committees, different test item writers and review committees, and so on).

Archbald (1994) conducted a study of four states' development of their respective curriculum guides. Here are some of his comments about the similarities across states:

> "Among the three states that include rationales in their state guides (California, Texas, and New York), there is considerable variation in how they address their purposes." (p.9)
>
> "… the state guides vary tremendously in how specifically topics are described." (p.18)
>
> "There is no single formula for the format, organization, or detail of state curriculum guides. The great variation in the rationales and prescriptiveness of the states' guides testifies to the lack of consensus concerning their optimal design." (p.21)

In a study contrasting the wide variety of different district responses in standards development to state standards initiatives, Massell, Kirst, and Hoppe (1997, p.7) wrote:

> "… most of the districts in our sample were actively pursuing their own standards-based curricular and instructional change. While state policies often influenced local efforts in this direction, it is important to note that many districts led or substantially elaborated upon state initiatives.
>
> "Rather than stunting local initiative and decisionmaking, state action could stimulate, but it did not uniformly determine, districts' and schools' own curricular and instructional activities.
>
> "… local staff in nearly all the sites typically regarded the state's standards as only one of many resources they used to generate their own, more detailed curricular guidance policies and programs. They reported turning to multiple sources—the state, but also to national standards groups, other districts, and their own communities—for input to develop their own, tailored guidance documents."

Buckendahl, Plake, Impara, and Irwin (2000) compared the test/standards alignment processes of test publishers for two test batteries that were also, and separately, aligned by panels of teachers. The comparison revealed inconsistencies:

> "The results varied across the two tests and the three grade levels. For example, the publisher indicated that 11 of the reading/language arts standards at grade 4 were aligned with Test A. The panel of teachers found only six of these standards aligned with this test (a 55% agreement). For Test B, the discrepancy was even greater. The publisher found that 14 of the 16 standards were assessed and the teachers found only six of the standards to be aligned (a 43% agreement)." (Bhola, Impara, & Buckendahl 2003, p. 28)

Given all this variety, why should anyone expect two different, separately-developed tests

in the same subject area to "generalize" to each other?

Over the past dozen years, state and local curricular standards for mathematics have probably become more similar than they were in 1990, thanks to the standardizing influence of the *Principles and Standards for School Mathematics* (1991) of the National Council of Teachers of Mathematics (NCTM), the main professional association of elementary and secondary mathematics teachers. The first edition of the NCTM Standards did not appear until the early 1990s. Even with the homogenous influence of a common, and widely available, set of mathematics standards, though, one can still find substantial differences from state to state, easily enough to account for the difference in average achievement test scores claimed by the CRESST researchers (which was one half a grade-level equivalence). Besides, the early editions of the NCTM Standards did less to set what mathematics should be learned than to set forth a general approach to teaching mathematics.

I performed a simple Web search on primary grades state mathematics standards and downloaded those from the first four states showing in the resulting list. Those states are Arizona, California, North Carolina, and Tennessee. I turned first to content standards for "data analysis and probability"—a topic likely not even included in most U.S. primary grades prior to 1990. Within this topic, there are many similarities to what these four states expect their students to know and be able to do by third grade. But, there also are substantial differences, differences that surely manifest themselves in what the students are taught and also in what gets included in their third-grade tests.

In Exhibit 2, I list just some of the topics, within just one of several strands of standards within mathematics that can be found either in one state's standards, or in two states' standards, but not in the other states' standards. Multiply the number of topics listed in Exhibit 2 by tenfold, and one still would not arrive at the number of discrepancies in content standards across just these four states, in just one subject area, at just one level of education. Then, ask yourself why a third grade student in Tennessee should be able to perform just as well on a third grade mathematics test in Arizona as on a Tennessee third grade mathematics test.

Exhibit 2.

Here are just some of the standards that exist for one of the four states, but not those for any of the three others, by (STATE, grade level):

"collect and record data from surveys (e.g., favorite color or food, weight, ages…" (AZ, 1)
"identify largest, smallest, most often recorded (i.e., mode), least often and middle (i.e., median) using sorted data" (AZ, 3)
"formulate questions from organized data" (AZ, 3)
"answer questions about a circle graph (i.e., pie graph) divided into 1/2s and 1/4s" (AZ, 3)
"answer questions about a pictograph where each symbol represents multiple units" (AZ, 3)
"write a title representing the main idea of a graph" (AZ, 3)
"locate points on a line graph (grid) using ordered pairs" (AZ, 3)
"predict the most likely or least likely outcome in probability experiments" (AZ, 3)
"compare the outcome of the experiment to the predictions" (AZ, 3)
"identify, describe, and extend simple patterns (such as circles or triangles) by referring to their shapes, sizes, or colors)" (CA, K)
"describe, extend, and explain ways to get to a next element in simple repeating patterns (e.g., rhythmic, numeric, color, and shape)" (CA, 2)
"sort objects and data by common attributes and describe the categories" (CA, 2)
"identify features of data sets (range and mode)" (CA, 2)
"determine the number of permutations and combinations of up to three items" (NC, 3)
"solve probability problems using permutations and combinations" (NC, 3)
"collect, organize, describe and display data using Venn diagrams (three sets) and pictographs where symbols represent multiple units (2., 5s, and 10s)" (NC, 2)
"collect and organize data as a group activity" (NC, K)
"display and describe data with concrete and pictorial graphs as a group activity" (NC, K)

Here are just some of the standards that exist for two of the four states, but not for the other two, by (STATE, grade level)

"collect and record data from a probability experiment" (AZ, 3)(CA,3)
"identify whether common events are certain, likely, unlikely, or improbably" (CA, 3) (TN, 2)
"ask and answer simple questions related to data representations" (CA, 2)(TN, 2)
"represent and compare data (e.g., largest, smallest, most often, least often) by using pictures, bar graphs, tally charts, and picture graphs)" (CA, 1)(TN, 1)
"use the results of probability experiments to predict future events (e.g., use a line plot to predict the temperature forecast for the next day) (CA, 3)(NC, 2)
"make conjectures based on data gathered and displayed" (TN, 3)(AZ, 2)
"pose questions and gather data to answer the questions (TN, 2)(CA, 2)
"conduct simple probability experiments, describe the results and make predictions (NC, 2) (AZ, 2)

SOURCES: Arizona Department of Education; California Department of Education; Hubbard; North Carolina Department of Public Instruction.

More LOVB: Title I Testing and the Lost Summer Vacation

> This tendency for scores to be inflated and therefore give a distorted impression of the effectiveness of an educational intervention is not unique to TIERS. Nor is it only of historical interest.
> – R.L. Linn, CRESST 2000, p.5

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, p.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 1991) 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 of all the factors 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.

Why Low Stakes are Associated with Test Score Inflation

> When high stakes kick in, the lack of public-ness and of explicitness of test attributes, lead teachers, school personnel, parents, and students to focus on just one thing: raising the test score by any means necessary.
> – E.L. Baker, CRESST 2000

Given current law and practice, the typical high-stakes test is virtually certain to be accompanied by item rotation, sealed packets, monitoring by external proctors, and the other test security measures itemized as necessary by Cannell in his late-1980s appeal to clean up the rampant corruption in educational testing and reporting.[17]

Two decades ago, Cannell suspected a combination of educator dishonesty and lax test security to be causing test score inflation. But, educators are human, and educator dishonesty (in at least some proportion of the educator population) is not going away any time soon. So, if Cannell's suspicions were correct, the only sure way to prevent test score inflation would be with tight test security. In Cannell's review of 50 states and even more tests, testing programs with tight security had no problems with test score inflation.

High-stakes are associated with reliable test results, then, because high-stakes tests are administered under conditions of tight test security. That security may not always be as tight as it could be, and may not always be as tight as it should be, but it is virtually certain to be much tighter than the test security that accompanies low- or no-stakes tests (that is, when the low- or no-stakes tests impose any test security at all).

In addition to current law and professional practice, other factors that can enhance test

---

[17] Most of the procedures Cannell recommended can be found in the 1999 *Standards*: specifically among standards 1.1–1.24 (test validity), pp. 9–24; standards 3.1–3.27 (test development and revising), pp. 37–48; standards 5.1–5.16 (test administration, scoring, and reporting), pp.61–66; and standards 8.1–8.13 (rights and responsibilities of test takers), pp. 85–90.

security, that also tend to accompany high stakes tests, are a high public profile, media attention, and voluntary insider (be they student, parent, or educator) surveillance and reporting of cheating. Do a Web search of stories of test cheating, and you will find that, in many cases, cheating teachers were turned in by colleagues, students, or parents. (See, for example, the excerpts from "Cheating in the News" at www.caveon.com.)

Public attention does not induce otherwise honest educators to cheat, as the researchers at the Center for Research on Education Standards and Student Testing (CRESST) claim. The public attention enables otherwise successful cheaters to be caught. In contrast to Baker's (2000) assertion quoted above, under current law and practice, it is typically *high-stakes* tests that are public, transparent, and explicit in their test attributes and public objectives, and it is typically low-stakes tests that are not.

Conclusion

People only know what you tell them.
– Frank Abagnale, Jr.[18]

What happens to the virtuous teachers and administrators in Lake Wobegon who vigorously maintain moral standards in the midst of widespread cheating? Those with the most intrepid moral characters risk being classified as the poorer teachers after the test scores are summarized and posted—with their relatively low, but honest scores compared to their cheating colleagues' inflated, but much higher scores.

Likewise, any new superintendent hired into a school district after a several-year run-up in scores from a test score pyramid scheme faces three choices—administer tests honestly and face the fallout from the resulting plunge in scores; continue the sleight-of-hand in some fashion; or declare standardized tests to be invalid measures of "real learning," or some such, and discontinue the testing. There are few incentives in Lake Wobegon to do the right thing.

The Cannell Reports remain our country's most compelling and enduring indictment of education system self-evaluation. But, most education research assumes that educators are incapable of dishonesty, unless unreasonably forced to be so. So long as mainstream education research demands that educators always be portrayed as morally beyond reproach, much education research will continue to be stunted, slanted, and misleading.

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;

[18] Attributed to con man Frank Abagnale, Jr., as played by Leonardo DiCaprio, in *Catch Me If You Can*, Dreamworks Productions, LLC, 2002.

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.

Table 10 summarizes the test-score inflation dynamic succinctly.

Table 10.

	Test security is TIGHT	Test security is LAX
Stakes are HIGH	no	yes
Stakes are LOW	no	yes

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.

Postscript 1. Yet More Left-Out-Variable-Bias (LOVB)

"Schools have no incentive to manipulate scores on these nationally respected tests…."
J.P. Greene, et al. 2003, Executive Summary

Illustrating the wide spread of the belief in the high-stakes-cause-test-score-inflation hypothesis, even some testing advocates have accepted it as correct. Read for example, the statement above by Jay P. Greene of the Manhattan Institute.

If you assume that he must be referring to valid, high-stakes standards-based tests, you would be assuming wrongly. He is referring, instead, to national norm-referenced tests (NRTs) taken "off-the-shelf" and then, perhaps, used legitimately as highly informative diagnostic instruments under conditions of high security or, perhaps, administered under who-knows-what conditions of test security and used to manufacture artificial test score gains. He is calling the type of test sometimes used in Lake Wobegon "nationally respected" and un-manipulated.

The Manhattan Institute's Work

Here's what Greene and his associates did. They gathered average test score data from two states and several large school districts. The jurisdictions they chose were special in that they administered both high-stakes standards-based tests and low- or no-stakes NRTs systemwide. They calculated standard correlation coefficients between student high-stakes test scores and student low-stakes test scores. In a few cases the same students took both tests but, more often, the two tests were taken by two different groups of students from nearby grades, but still in the same jurisdiction. They also calculated standard correlation coefficients for gain scores (over years with synthetic cohorts) between high- and low-stakes test scores. (Greene, Winters, & Forster 2004)

Greene, et al, claim to have controlled for background demographic factors, as they only compared scores from the same jurisdiction. But, they did nothing to control for degrees of difference in the stakes and, more to the point, they did nothing to control for variations in test security or curricular content. Indeed, they declared the curricular content issue irrelevant (2003 pp.5, 6).

> "There is no reason to believe that the set of skills students should be expected to acquire in a particular school system would differ dramatically from the skills covered by nationally-respected standardized tests. Students in Virginia need to be able to perform arithmetic and understand what they read just like students in other places, especially if students in Virginia hope to attend colleges or find employment in other places."

Whether or not content standards *should* or *should not* differ dramatically across jurisdictions is irrelevant to the issue. The fact is that they can and they do. (see, for example, Archbald 1994; Massell, Kirst, & Hoppe 1997) Talk to testing experts who have conducted standards or curricular match studies, and one will learn that it is far from unusual for a nationally-standardized NRT to match a state's content standards at less than 50 percent. Such a low rate of match would suggest that more than half of the NRT items test content to which the state's students probably have not been exposed, more than half of the state's content standards are not tested by the NRT, or some combination of both.

In sum, the Manhattan Institute report concurs with testing critics' assertions that:

externally-administered high-stakes testing causes score inflation, and that internally-administered low- or no-stakes testing does not;

"teaching to the test" (which occurs naturally with good alignment) is a bad thing, and is measurable; and

it is legitimate to measure the "real" score increases of high-stakes standards-based tests only with an unrelated low-stakes shadow test, regardless of the curricular content match, or lack thereof, between the two tests.

Manhattan Institute Says Incentives Don't Matter

Furthermore, the Manhattan Institute report concurs with the suggestion of the Center for Research on Education Standards and Student Testing (CRESST) that there is no correlation between high-stakes, increases in motivation, and increases in achievement, in the manner explained below.

Controlled experiments from the 1960s through the 1980s tested the hypothesis (see Phelps 2005, Appendix B). Half of the students in a population were assigned to a course of study and told there would be a final exam with consequences (reward or punishment) riding on the results. The other half were assigned to the same course of study and told that their performance on the final exam would have no consequences. Generally, there were no incentives or consequences for the teachers. Guess which group of students studied harder and learned more?

The Manhattan Institute has apparently joined with CRESST in ruling out the possibility of motivation-induced achievement gains. With their methodology, any increase in scores on a high-stakes test exceeding increases in an unrelated parallel no-stakes test must be caused by "teaching to the test," and is, thus, an artificial and inflated score gain ...not evidence of "real learning."

Unreliable Results

Still another irony is contained in the Greene et al., claim that NRTs are "nationally respected tests" and the quality of state standards-based tests can be judged by their degree of correlation with them. They calculated, for example, a 0.96 correlation coefficient between Florida's high stakes state test and a low-stakes NRT used in Florida. (Greene, et al., Executive Summary) This degree of correlation would be considered high even for two forms of the same test.[19]

By contrast, Greene et al. calculated a 0.35 correlation coefficient between Colorado's high-stakes state test and a low-stakes NRT used in Fountain Fort Carson, CO. (Greene, et al., Executive Summary) This is a remarkably low correlation for two tests claiming to measure

[19] As Taylor (2002, p.482) put it:

"If two tests are supposed to measure exactly the same content and skills (for example, two forms of the Iowa Test of Educational Development (ITED)), the correlations should be *very high* (about .90)

"If two tests are supposed to measure similar knowledge and skills but also have differences in terms of the targeted knowledge and skills, the correlations should be *strong* but *not too high* (between .70 and .80)."

achievement of similar subject matter. So, to borrow the authors' words, one cannot "believe the results of" the accountability test in Colorado or, at least those in Fountain Fort Carson, CO? I would strongly encourage anyone in Fountain Fort Carson, CO. to first consider the left out variables—variation in curricular content covered, variation in the degree of test security, and others—before jumping to that conclusion.

Any state following the Greene, et al. logic should prefer to have their high-stakes standards-based tests developed by the same testing company from which they purchase their low-stakes NRTs. Likewise, any state should eschew developing their high-stakes tests independently, in an effort to maximize the alignment to their own curriculum. Moreover, any state should avoid custom test-development processes that involve educators in writing or reviewing standards, frameworks, and test items because the more customized the test, the lower the correlation is likely to be with the off-the-shelf NRTs.

In other words, the tighter the alignment between a jurisdiction's standards-based test and its written and enacted curriculum, the lower the quality of the test… at least according to the Manhattan Institute.

Postscript 2. Actually, Tests Do Not Talk

"Score inflation is a preoccupation of mine."
Daniel Koretz
Measuring Up: What Educational Testing Really Tells Us, 2008

In his 2008 book, *Measuring Up*, Daniel Koretz continues his defense of the theory with which he is most famously identified: He argues that high stakes induce "teaching to the test," which in turn produces artificial test-score gains (i.e., test-score inflation). The result, according to Koretz:

> "Scores on high-stakes tests-tests that have serious consequences for students or teachers-often become severely inflated. That is, gains in scores on these tests are often far larger than true gains in students' learning. Worse, this inflation is highly variable and unpredictable, so one cannot tell which school's scores are inflated and which are legitimate." (p. 131)

Thus, Koretz, a long-time associate of the federally funded Center for Research on Educational Standards and Student Testing (CRESST), provides the many educators predisposed to dislike high-stakes tests anyway a seemingly scientific (and seemingly not self-serving or ideological) argument for opposing them. Meanwhile, he provides policymakers a conundrum: if scores on high-stakes tests improve, likely they are meaningless-leaving them no external measure for school improvement. So they might just as well do nothing as bother doing anything.

Measuring Up supports this theory by ridiculing straw men—declaring a pittance of flawed supporting evidence sufficient (pp. 11, 59, 63, 132, & chapter 10) and a superabundance of contrary evidence nonexistent-and mostly by repeatedly insisting that he is right.(See, for example, chapter 1, pp. 131–133, & 231-236.) He also shows little patience for those who choose to disagree with him. They want "simple answers", speak "nonsense", assert "hogwash", employ "logical sleight(s) of hand", write "polemics", or are "social scientists who ought to know better".

Lake Wobegon

The concept of test-score inflation emerged in the late 1980s from the celebrated studies of the physician John J. Cannell (1987, 1989). Dr. Cannell caught every U.S. state bragging that its students' average scores on national norm-referenced tests were "above the national average," a mathematical impossibility. The phenomenon was dubbed the "Lake Wobegon Effect," in tribute to the mythical radio comedy community in which "all the children are above average."

What had caused the Lake Wobegon Effect? Cannell identified several suspects, including educator dishonesty and conflict of interest; lax test security; and inadequate or outdated norms. But Cannell's seemingly straightforward conclusions did not make it unscathed into the educational literature. For instance, one prominent CRESST study provided a table with a cross-tabulation that summarized (allegedly all) the explanations provided for the spuriously high scores (Shepard 1990, 16). Conspicuously absent from the table, however, were Cannell's two primary suspects-educator dishonesty and lax test security.

Likewise, Koretz and several CRESST colleagues followed up with their own study in an unnamed school district, with unnamed tests and unidentified content frameworks. Contrasting a steadily increasing rise in scores on a new, "high stakes" test with the substantially lower scores recorded on an older, no-stakes test, Koretz and his colleagues attributed the inflation to the alleged high stakes.[20] Not examined was why two different tests, developed by two completely different groups of people under entirely separate conditions, using no common standard for content, would be expected to produce nearly identical scores.

This research framework presaged what was to come. The Lake Wobegon Effect continued to receive considerable attention, but Cannell's main points-that educator cheating was rampant and test security inadequate-were dismissed out of hand and persistently ignored thereafter. The educational consensus, supported by the work of CRESST and other researchers, fingered "teaching to the test" for the crime, manifestly under pressure from the high stakes of the tests.

Problematically, however, only one of Cannell's dozens of score-inflated tests had any stakes attached. All but that one were no-stakes diagnostic tests, administered without test-security protocols. The absence of security allowed education administrators to manipulate various aspects of the tests' administration, artificially inflate scores, and then advertise the phony score trends as evidence of their own managerial prowess. Ironically, many of the same states simultaneously administered separate, genuinely high-stakes tests with tight security and no evidence of score inflation.

Much of *Measuring Up* recapitulates the author's earlier writings, but on page 243, we do learn what he and his colleagues actually found in that influential follow-up to Cannell's findings. Exactly why had scores risen so dramatically on the new, high-stakes third-grade test they examined?

> "[A]lthough the testing system in this district was considered high-stakes by the standards of the late 1980s, by today's standards it was tame. There were no cash awards . . . threats to dissolve schools or remove students in response to low scores. . . . The pressure arose only from less tangible things, such as publicity and jawboning."

In other words, this foundational study had involved no real high-stakes test at all. After all,

---

[20] The study traced the annual trend in average scores on a third-grade test "perceived to be high stakes" over several years, then administered a different third-grade test, with no stakes, that had been administered in the district several years earlier. The researchers, finding a steadily increasing rise in scores on the new test contrasted with a substantially lower score on the old, no-stakes test, attributed the rise in scores on the new test to inflation allegedly caused by the alleged high stakes. The study ignored several factors that could have influenced the results, such as differing content, teachers, students, and incentives. Indeed, it ignored most of the factors that could have influenced the results, or speculated that they must have conveniently cancelled each other out, and then declared that high stakes must have done it.

Even nearly two decades later, much of the study remains shrouded in mystery: "The price of admission [to conduct the study] was that we take extraordinary steps to protect the anonymity of the [school] district, so I cannot tell you its name, the state it was in, or even the names of the tests we used." Thus, the study is neither replicable nor falsifiable. An easy solution would be a content match study between the two tests used for comparison. If, as claimed, the two tests represented the same domain (identified, i.e., it could have been [and likely was] as broad as a "grade level" of mathematics from two completely different content frameworks with non-parallel topical sequences), why not support that assertion with some empirical evidence?

in our open democracy, all tests are subject to "publicity and jawboning," whether they genuinely have high stakes or no stakes. (Koretz, incidentally, is also incorrect in characterizing the test as "high stakes by the standards of the late 1980s": at the time more than twenty states administered high school graduation exams—for which failing students were denied diplomas.)

Do as I Say, Not as I Do

Many testing researchers (unsurprisingly, not associated with CRESST) caution against the simplistic assumptions that any test will generalize to any other simply because they have the same subject field name or that one test can be used to benchmark trends in the scores of another (Archbald, 1994; Bhola, Impara, and Buckendahl, 2003, 28; Buckendahl, et al., 2000; Cohen and Spillane, 1993, 53; Freeman, et al., 1983; Impara, 2001; Impara, et al., 2000; Plake, et al., 2000). Ironically, despite himself, Koretz cannot help agreeing with them. Much of the space in *Measuring Up* is devoted to cautioning the reader against doing exactly what he does-making apples-to-oranges comparisons with scores or score trends from different tests. For example:

> “One sometimes disquieting consequence of the incompleteness of tests is that different tests often provide somewhat inconsistent results.” (p. 10)
>
> “Even a single test can provide varying results. Just as polls have a margin of error, so do achievement tests. Students who take more than one form of a test typically obtain different scores.” (p. 11)
>
> “Even well-designed tests will often provide substantially different views of trends because of differences in content and other aspects of the tests' design. . . . [W]e have to be careful not to place too much confidence in detailed findings, such as the precise size of changes over time or of differences between groups.” (p. 92)
>
> “[O]ne cannot give all the credit or blame to one factor . . . without investigating the impact of others. Many of the complex statistical models used in economics, sociology, epidemiology, and other sciences are efforts to take into account (or 'control' for') other factors that offer plausible alternative explanations of the observed data, and many apportion variation in the outcome-say, test scores-among various possible causes. …A hypothesis is only scientifically credible when the evidence gathered has ruled out plausible alternative explanations.” (pp. 122-123)
>
> “[A] simple correlation need not indicate that one of the factors causes the other.” (p. 123)
>
> “Any number of studies have shown the complexity of the non-educational factors that can affect achievement and test scores.” (p. 129)

Recommendation recoil

Koretz's vague suggestion that educators teach to "a broader domain" would dilute coverage of required content that typically has been developed through a painstaking public process of expert review and evaluation. In its place, educators would teach what exactly? Content that Koretz and other anti-standards educators prefer? When the content domain of a test is the legally (or intellectually) mandated curriculum, teachers who "teach to the test" are

not only teaching what they are told they should be teaching, they are teaching what they are legally and ethically obligated to teach (Gardner 2008).

Another example of an imprudent recommendation: the Princeton Review sells test preparation services, most prominently for the ACT and SAT college admission tests. Its publishers argue that students need not learn subject matter to do well on the tests, only learn some test-taking tricks. Pay a small fortune for one of their prep courses and you, too, can learn these tricks, they advertise. Curiously, independent studies have been unable to confirm Review's claims (see, for example, Camara, 2008; Crocker, 2005; Palmer, 2002; Tuckman, 1994; Tuckman and Trimble, 1997; Allensworth, Correa, & Ponisciak, 2008), but Koretz supports them: "…this technique does often help to raise scores."

After investigations and sustained pressure from better business groups, the Princeton Review in 2010 voluntary agreed to pull its advertising that promised score increases from taking its courses (National Advertising Division, 2010).

Scripting a hoax

Around 1910, a laborer at the Piltdown quarries of southern England discovered the first of two skulls that appeared to represent the missing link between ape and human. In the decades following, mainstream science and some of the world's most celebrated scientists would accept "Piltdown Man" as an authentic specimen of an early hominid. Along the way, other scientists, typically of the less famous variety, proffered criticisms of the evidence, but were routinely ignored. Only in the 1950s, after a new dating technique applied to the fossil remains found them to be modern, was the accumulated abundance of contrary evidence widely considered. The Piltdown fossils, it turned out, were cleverly disguised forgeries.

"Piltdown man is one of the most famous frauds in the history of science," writes Richard Harter in his review of the hoax literature (1996-1997). Why was it so successful? Harter offers these explanations:

- some of the world's most celebrated scientists supported it;
- it matched what prevailing theories at the time had led scientists to expect;
- various officials responsible for verification turned a blind eye;
- the forgers were knowledgeable and skilled in the art of deception;
- the evidence was accepted as sufficient despite an absence of critical details; and
- contrary evidence was repeatedly ignored or dismissed.

Measuring Up's high-stakes-cause-test-score-inflation myth-making fits the hoax script perfectly.

REFERENCES

Allensworth E., Correa, M., Ponisciak, S. (2008, May). *From High School to the Future: ACT Preparation–Too Much, Too Late: Why ACT Scores Are Low in Chicago and What It Means for Schools*. Chicago, IL: Consortium on Chicago School Research at the University of Chicago.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Federation of Teachers. 1995. *Defining World Class Standards*. Washington, D.C.

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, April.

Arizona Department of Education. *Mathematics Standards Chart for AIMS: Standards 1 through 6: Foundations level (Grade 3)*. [downloaded 09/17/05, from www.ade.state.az.us/standards/contentstandards.htm]

Baker, E.L. (2000). *Understanding educational quality: Where validity meets technology*. William H. Angoff Memorial Lecture. Educational Testing Service, Policy Information Center, Princeton, NJ.

Bangert-Drowns, R.L., J.A. Kulik, and C-L.C. Kulik. 1991.Effects of frequent classroom testing, *Journal of Educational Research*. 85(1), November/December, pp.89–99.

Barton, P.E. (1999a). Tests, tests, tests: We are giving too many tests—and they are not telling us what we need to know. *American Educator*, Summer [available online at www.aft.org].

Barton, Paul E. (1999b). *Too Much Testing of the Wrong Kind: Too Little of the Right Kind in K-12 Education*, Policy Information Center, Educational Testing Service, March.

Becker, B.J. 1990. Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal, *Review of Educational Research,* 60(3), Fall, 373–417.

Bhola, D.D., Impara, J.C., & Buckendahl, C.W. (2003). Aligning tests with States' content standards: Methods and issues, *Educational Measurement: Issues and Practice*. Fall, pp.21–29.

Bishop, J.H. (1997). Do curriculum-based external exit exam systems enhance student achievement? Paper #97-28. Center for Advanced Human Resource Studies, Institute for Labor Relations, Cornell University, Ithaca, November.

Bracey, G.W. (2000). The 10th Bracey Report un the condition of public education. *Phi Delta Kappan*. October, 133–144.

Briggs, D.C. (2001). The effect of admissions test preparation. *Chance*, Winter.

Briggs, D. & B. Hansen. (2004). Evaluating SAT test preparation: Gains, effects, and self-selection. Paper presented at the Educational Testing Service, Princeton, NJ, May.

Britton, E.D. & Raizen, S.A. (1996). *Examining the Examinations: An International Comparison of Science and Mathematics Examinations for College-Bound Students*. Boston: 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.

Buckendahl, C.W & Hunt, R. (2005). Whose rules? The relation between the "rules" and "law" of testing. In R.P. Phelps (Ed.), *Defending Standardized Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, pp.147–158.

California Department of Education. *Content Standards*. [downloaded 09/17/05, from www.cde.ca.gov/be/st/ss/mthgrade3.htm] & [...mthgrade2.htm] & […mthgrade1.htm] & […mthgradek.htm]

Camara, W. (1999). Is commercial coaching for the SAT I worth the money? *College Counseling Connections*. The College Board. New York, NY, 1(1), Fall.

Camara, W. J. 2008. College Admission Testing: Myths and Realities in an Age of Admissions Hype, in *Correcting Fallacies about Educational and Psychological Testing* (chapter 4), ed. R.P. Phelps. Washington, D.C.: 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*. Daniels, WV: Friends for Education.

Cannell, J.J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston: Allyn & Bacon.

Cheng, L. & Y. Watanabe. (2004). *Washback in language testing: Research contexts and methods*. Lawrence Erlbaum Associates. Mahwah, NJ.

Cohen, D.K. & Spillane, J.P. (1993). Policy and practice: The relations between governance and instruction. In Fuhrman, S.H., Ed. *Designing Coherent Education Policy: Improving the System*. San Francisco: Jossey-Bass, 35–95.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*. 66(3), Fall.

Crocker, L. (2005). Teaching for the Test: How and Why Test Preparation Is Appropriate. In *Defending Standardized Testing*, ed. R. P. Phelps, 159-174. Mahwah, N.J.: Lawrence

Erlbaum.

Debra P. v. Turlington, 644 F.2d 397, 6775 (5th Cir. 1981).

DerSimonian and Laird, 1983. Evaluating the effect of coaching on SAT scores: A meta-analysis, *Harvard Educational Review* 53, 1-5.

Eckstein, M.A. & Noah, H.J. (1993). *Secondary School Examinations: International Perspectives on Policies and Practice*. New Haven: Yale University Press.

Fraker, G.A. (1986–87). The Princeton Review reviewed. *The Newsletter*. Deerfield, MA: Deerfield Academy, Winter.

Fredericksen, N. (1994). *The Influence of Minimum Competency Tests on Teaching and Learning*. Princeton, NJ: Educational Testing Service.

Freeman, D., and others. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *Elementary School Journal*. Vol.83, No.5, 501–514.

Gardner, W. 2008, April 17. "Good Teachers Teach to the Test: That's Because It's Eminently Sound Pedagogy." Christian Science Monitor.

GI Forum et al. v. Texas Education Agency et al., F.Supp, 1 (W.D. Tex. 2000)

Harter, R. 1996-1997. Piltdown Man: The Bogus Bones Caper. The TalkOrigins Archive. Downloaded May 13, 2008, from http://www.talkorigins.org/faqs/piltdown.html.

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, Washington.

Impara, J. C., B. S., Plake, and C. W. Buckendahl. 2000, June. 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, Utah.

Greene, J.P., Winters, M.A., & Forster, G. (2003). *Testing high-stakes tests: Can we believe the results of accountability tests?* Manhattan Institute, Center for Civic Innovation, Report #33.

Greene, J.P., Winters, M.A., & Forster, G. (2004). Testing high-stakes tests: Can we believe the results of accountability tests? *Teachers College Record*, 106(6), June, pp.1124–1144.

Hartigan, J.A. & Wigdor, A.K. (1989). *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery.* Washington, D.C.: National Academy Press.

Hatch, J.A. & Freeman, E.B. (1988). Who's pushing whom? Stress and kindergarten. *Phi Delta Kappan*, 69, 145–147.

Heubert, J.P. & Hauser, R.M. (1999). *High Stakes: Testing for Tracking, Promotion, and*

Graduation. Washington, D.C.: National Academy Press.

Hubbard, I. *Mathematics curriculum standards: Kindergarten–Third Grade*. (Approved by the Tennessee State Board of Education) [downloaded 0917/05, from www.state.tn.us/education/ci/cistandards2001/math/cimathk3stand.htm]

Koretz, D. (1992). NAEP and the movement toward national testing. Paper presented in Sharon Johnson-Lewis (Chair), *Educational Assessment: Are the Politicians Winning?* Symposium presented at the annual meeting of the American Educational Research Association, San Francisco, April 22.

Koretz, D.M. (1996). Using student assessments for educational accountability, in E.A. Hanushek & D.W. Jorgenson, Eds. *Improving America's schools: The role of incentives*. Washington, D.C.: National Academy Press.

Koretz, D.M. (2008). *Measuring up: What educational testing really tells us.* Harvard University Press, 2008.

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991) The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented in 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, April 5.

Kulik, J.A., Bangert-Drowns, R.L. and C-L.C. Kulik 1984. "Effectiveness of coaching for aptitude tests," *Psychological Bulletin 95*, 179-188.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *CSE Technical Report 351*. Center for Research on Education Standards and Student Testing, January.

Linn, R.L. (1995). Assessment-based reform: Challenges to educational measurement. William H. Angoff Memorial Lecture. Educational Testing Service, Princeton, NJ.

Linn, R.L. (1998). Standards-based accountability: Ten suggestions. *CRESST Policy Brief*, adapted from CRESST Technical Report 490, *Standards-Led Assessment: Technical and Policy Issues in Measuring School and Student Progress*. Los Angeles.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher*. March, 4–16.

Linn, R.L., Dunbar, S.B., Harnisch, D.L. & Hastings, D.N. (1982). The validity of the Title I evaluation and reporting system. In E.R. House, S. Mathison, J. Pearsol, & H. Preskill (Eds.), *Evaluation studies review annual* (vol.7, pp.427–442). Beverly Hills, CA: Sage Publications.

Linn, R.L., Graue, M.E., & N.M. Sanders. (1990). Comparing state and district results to national norms: The validity of the claims that 'everyone is above average.' *Educational Measurement: Issues and Practice*, 9(3), 5–14.

Massell, D., Kirst, M. & Hoppe, M. (1997). Persistence *and change: Standards-based systemic reform in nine states*. CPRE Policy Brief, RB-21, Consortium for Policy Research in Education, March.

McNeil, L.M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

McNeil, L. and A. Valenzuela (2000). *The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric*, The Civil Rights Project, Harvard University,

Messick, S. & A. Jungeblut (1981). Time and method in coaching for the SAT, *Psychological Bulletin* 89, 191–216.

National Advertising Division. (2010). The Princeton Review voluntarily discontinues certain advertising claims; NAD finds company's action 'necessary and appropriate'. New York, NY: National Advertising Review Council, Council of Better Business Bureaus, CBBB Children's Advertising Review Unit, National Advertising Review Board, and Electronic Retailing Self-Regulation Program. Retrieved June 25, 2010 from http://www.nadreview.org/DocView.aspx?DocumentID=8017&DocType=1

National Center for Education Statistics (NCES) (1993). *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States*. Report #23–ST04. Washington, DC: May.

National Council of Teachers of Mathematics (1991). *Principles and Standards for School Mathematics*. Author. Washington, D.C.

North Carolina Department of Public Instruction. NC Standard Course of Study: Mathematics. [downloaded 09/17/05, from www.ncpublicschools.org/curriculum/mathematics/scos/2003/k-8/16grade2.htm]

NEAToday Online (2002). Interview with Lorrie Shepard: How to fight a 'Death Star' January 2001.

Palmer, J. S. 2002. Performance Incentives, Teachers, and Students: Estimating the Effects of Rewards Policies on Classroom Practices and Student Performance. PhD dissertation. Columbus, Ohio: Ohio State University.

Phelps, R.P. (1996). Are U.S. students the most heavily tested on earth? *Educational Measurement: Issues and Practice*, Vol.15, No.3, Fall, 19–27.

Phelps, R.P. (1999). Education establishment bias? A look at the National Research Council's critique of test utility studies. *The Industrial-Organizational Psychologist*. 36(4), April, 37–49.

Phelps, R.P. (2000). Trends in large-scale, external testing outside the United States. *Educational Measurement: Issues and Practice*. Vol.19, No.1, 11–21.

Phelps, R.P. (2001). Benchmarking to the world's best in mathematics: Quality control in curriculum and instruction among the top performers in the TIMSS. *Evaluation Review*, Vol.25, No.4, August, 391–439.

Phelps, R.P. (2005). The rich, robust research literature on testing's achievement benefits, in

R.P. Phelps, Ed. *Defending Standardized Testing*. Lawrence Erlbaum Associates. Mahwah, NJ, 55–90.

Phillips, G.W. & Finn, C.E., Jr. (1988). The Lake Wobegon Effect: A skeleton in the testing closet? *Educational Measurement: Issues and Practice*. Summer, 10–12.

Plake, B. S., C. W. Buckendahl, and J. C. Impara. 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.

Popham, W.J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, 68, 675–682.

Powers, D.E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice*. 24–30, 39.

Powers, D.E. & D.A. Rock. (1999). Effects of coaching on SAT I: Reasoning Test scores. *Journal of Educational Measurement*. 36(2), Summer, 93–118.

Robb, T.N. & J. Ercanbrack. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *Teaching English as a Second or Foreign Language*. v.3, n.4, January.

Shepard, L.A. (1989). Inflated test score gains: Is it old norms or teaching the test? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, March.

Shepard, L.A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*. Fall, 15–22.

Shepard, L.A. (2000). The role of assessment in a learning culture. Presidential Address presented at the annual meeting of the American Educational Research Association, New Orleans, April 26.

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*. Prepared for the U.S. Department of Education, Office of Policy and Planning, 1992.

Smith, M.L. (1991a). The Role of Testing in Elementary Schools, *CSE Technical Report 321*, Los Angeles, UCLA, May.

Smith, M.L. (1991b). Put to the Test: The Effects of External Testing on Teachers, *Educational Researcher*, 20(5), June.

Smith, M.L. (1991c). Meanings of Test Preparation, *American Educational Research Journal*, 28(3), Fall.

Smith, M.L. & C. Rottenberg (1991). Unintended Consequences of External Testing in Elementary Schools, *Educational Measurement: Issues and Practice*. Winter, 10–11.

Smyth, F.L. (1990). SAT coaching: What really happens to scores and how we are led to expect more. *The Journal of College Admissions*, 129, 7–16.

Snedecor, P.J. (1989). Coaching: Does it pay—revisited. *The Journal of College Admissions*. 125, 15–18.

Taylor, C.S. (2002). Evidence for the reliability and validity of scores from the Washington Assessment of Student Learning (WASL). Paper presented to the Washington Roundtable. December 11.

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., and S. Trimble. 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. Available from ERIC (ED418785).

Whitla, D.K. (1988). Coaching: Does it pay? Not for Harvard students. *The College Board Review*. 148, 32–35.

Winfield, L.F. (1990). School competency testing reforms and student achievement: Exploring a national perspective. *Education Evaluation and Policy Analysis*, 12(2), 157–173.

Zehr, M.A. (2001). Study: Test-preparation courses raise scores only slightly. *Education Week*. April 4.