Revista Iberoamericana de Evaluación Educativa, 2017, 10(1), 33-49.
https://doi.org/10.15366/riee2017.10.1.002

The "Teaching to the Test" Family of Fallacies

La Familia de Falacias "Enseñando para el Examen"

Richard P. Phelps *

University of Pennsylvania

This article explains the various meanings and ambiguities of the phrase "teaching to the test" (TttT), describes its history and use as a pejorative, and outlines the policy implications of the popular, but fallacious, belief that "high stakes" testing induces TttT which, in turn, produces "test score inflation" or artificial test score gains. The history starts with the infamous "Lake Wobegon Effect" test score scandal in the US in the 1980s. John J. Cannell, a medical doctor, discovered that all US states administering national norm-referenced tests claimed their students' average scores exceeded the national average, a mathematical impossibility. Cannell blamed educator cheating and lax security for the test score inflation, but education insiders managed to convince many that high stakes was the cause, despite the fact that Cannell's tests had no stakes. Elevating the high stakes causes TttT, which causes test score inflation fallacy to dogma has served to divert attention from the endemic lax security with "internally administered" tests that should have encouraged policy makers to require more external controls in test administrations. The fallacy is partly responsible for promoting the ruinous practice of test preparation drilling on test format and administering practice tests as a substitute for genuine subject matter preparation. Finally, promoters of the fallacy have encouraged the practice of "auditing" allegedly untrustworthy high-stakes test score trends with score trends from allegedly trustworthy low-stakes tests, despite an abundance of evidence that low-stakes test scores are far less reliable, largely due to student disinterest.

Keywords: Test security, Educator cheating, Test score inflation, High stakes, Standardized tests, Education.

Este artículo explica los diversos significados y ambigüedades de la frase "enseñar para el examen" (*TttT: teaching to the test* en inglés), describe su historia y su uso como un peyorativo, y describe las implicaciones políticas de la creencia popular, pero falaz, que las pruebas de a "gran escala" inducen TttT que, a su vez, produce una "inflación en la calificación obtenida en el examen" o ganancias em cuanto a los puntos obtenidos en la prueba. La historia comienza con el infame escándalo de la puntuación de la prueba "Lake Wobegon Effect" en los Estados Unidos en los años ochenta. John J. Cannell, un médico, descubrió que todos los estados de los Estados Unidos que administraban pruebas nacionales con referencias normativas afirmaban que los puntajes promedio de sus estudiantes excedían el promedio nacional, una imposibilidad matemática. Cannell atribuyó a los educadores el engaño y la seguridad laxa por la inflación de la puntuación de los exámenes, pero los expertos en educación lograron convencer a muchos de que las pruebas a gran escala eran la causa, a pesar de que las pruebas de Cannell no tenían ninguna fiabilidad. Exagerar las pruebas a gran escala hace que TttT hace que la falla de la inflación de la puntuación de la prueba al dogma haya servido para desviar la atención de la seguridad laxa endémica con pruebas "internamente administradas" que deberían haber alentado a los responsables políticos a exigir más controles externos en las administraciones de las pruebas. La falacia es en parte responsable de promover la práctica ruinosa en la preparación de las pruebas en el formato de prueba y la administración de pruebas prácticas como un sustituto de la preparación de la materia original. Por último, los promotores de la falacia han fomentado la

*Contacto: richardpphelps@yahoo.com

issn: 1989-0397
www.rinace.net/riee/
https://revistas.uam.es/riee

Recibido: 1 de octubre de 2016
1ª Evaluación: 3 de enero de 2017
Aceptado: 21 de febrero de 2017

práctica de "auditar" tendencias de determinadas puntuación en las pruebas a gran escala con las tendencias de puntuación presuntamente confiables de las pruebas de baja exigencia, a pesar de la abundancia de pruebas donde las puntuaciones de las pruebas a menor escala son mucho menos confiables debido al desinterés de los estudiantes.

Palabras clave: Prueba de seguridad, Engaño de educador, Inflación de la puntuación del examen, Pruebas a gran escala, Pruebas estandarizadas, Educación.

1. Introduction

Standardized testing is one of the few means by which the public may ascertain what transpires inside school classrooms and, by far, the most objective.

For those inside education who would prefer to be left alone to operate schools as they wish, externally managed standardized tests intrude. Many actively encourage public skepticism of those tests' validity. Promoting the concept of "teaching to the test" as a pejorative is one part of the effort (Phelps, 2011c).

But, the meaning of the phrase is ambiguous (Shepard ,1990; Popham, 2004). At worst, it suggests grossly lax test security: teachers know the exact contents of an upcoming test and expose their students to that content, thereby undermining the test as an objective measure. Some testing critics would have the public believe that this is always possible. It is not. When tests are secure, the exact contents are unknown to teachers and test-takers alike until the moment scheduled testing begins and they hear instructions such as "please break open the seal of your test booklet".

More often, the phrase "teaching to the test" (TttT) is used pejoratively when it allegedly induces teachers to reduce the quality of instruction. There are two ways this can happen.

First, TttT allegedly lowers educational quality due to the limitations of tests. Critics suggest that tests—or, typically, externally managed standardized tests—are not well correlated with learning. These tests cannot measure all that students learn, perhaps not even most of, or the best parts of, what they learn. If true, then teaching only those components of learning that tests can capture neglects other, allegedly important, components of learning.

For a skeptic, the assertion begs the question: if tests do not measure important components of learning, how do we know those components exist? The philosopher and mathematician René Descartes is said to have written, "If a thing exists, it exists in some amount. If it exists in some amount, it is capable of being measured". Was he wrong? Are there types of learning that teacher-made tests can capture, but standardized tests cannot? …that teachers can ascertain, but tests cannot? Is some learning simply immeasurable?

Most outside education probably assume that if a student cannot demonstrate a certain knowledge or skill on a test, that student probably does not possess that knowledge or skill.

Some inside education argue that standardized tests can only assess "lower-order skills" or "factual recall". So, teachers avoid more enlightening and challenging instruction in favor of the mundane and simple. Without tests, they argue, teachers teach and students learn higher and deeper knowledge and skills that cannot be validly assessed by

standardized tests. Rather, better knowledge and skills can only validly be assessed by methods that require a large amount of teacher observation and judgment. Long-term or group projects are sometimes mentioned as good vehicles for the demonstration of "better" student knowledge and skills.

The second way that "teaching to the test" allegedly lowers instructional quality is through test preparation. "Test prep" occurs in a variety of forms. The simplest form familiarizes test-takers with the structure and format of the test, and is unrelated to subject matter content. Format familiarization is particularly important when the format of an upcoming test is, …well… unfamiliar. If students, for example, have never seen a multiple-choice test item before, some instruction and practice can be helpful.

Opinions differ about how much instruction and practice is appropriate. Most testing experts and test developers advocate only a brief amount of time. How much time does it take to understand how to respond to a multiple-choice test item, after all? When a test format is so convoluted that extensive training is required to use it, format decoding may have become the skill being tested. Psychometricians would then say the test has "construct-irrelevant variance"-that is, it is testing skills and knowledge different from the intended "construct" (i.e., the subject matter content).

Many testing opponents and some test preparation companies, however, argue that extended practice (i.e., "drilling") on test format and practice tests can improve test performance (Fraker, 1986-87; Smyth, 1990; Marte, 2011). Unfortunately, some school personnel believe them and convert their classrooms into "test prep factories", halting regular subject-matter instruction in favor of instruction on standardized test formats, drilling with test-maker-provided workbooks, or administering practice tests (Shepard 1990).

All educators consider this type of TttT unfortunate and debilitating to learning. Educators disagree, however, as to whether it works to increase test scores.

Teaching to the test's negative connotations can befuddle naïve education outsiders who assume a natural complementary relationship between teacher instruction and student testing. Shouldn't teachers teach subject matter that will be on the test? Shouldn't a test include subject matter a teacher covered in class? Why would a teacher teach "away from the test"—deliberating teach subject matter that will not be tested or, conversely, test subject matter that was not taught?

If a test is aligned with subject-matter standards, and its questions thoroughly cover them, can responsible teachers avoid teaching to the test? (Gardner, 2008).

2. A short history of US educators cheating on tests

Teaching to the test (TttT) is far more than a catch phrase or slogan, however. It has served for three decades to divert attention from a more serious problem in education in the United States-educators cheating on assessments used to judge their own performance. To elaborate adequately requires a short history lesson first.

Arguably, the current prevalence of large-scale testing in the United States began in the late 1970s. Some statistical indicators revealed a substantial decline in student achievement from the early 1960s on. Many blamed perceived permissiveness and lowered standards induced by the social movements of the 1960s and 1970s. Statewide

testing-at least of the most basic skills-was proposed to monitor the situation. For motivation, some states added consequences to the tests, typically requiring a certain score for high school graduation.

With few exceptions (e.g., California, Iowa, New York), however, states had little recent experience in developing or administering standardized tests or writing statewide content standards. That activity had been deferred to schools and school districts. So, they chose the expedient of purchasing "off the shelf" tests-nationally norm-referenced tests (NRTs)[1] (Phelps 2008/2009a; 2010). Outside the states of Iowa or California, the subject matter content of NRTs matched that of no state. Rather, each covered a pastiche of content, a generic set thought to be fairly common.

Starting in the 1970s, the state of Florida required its high school students to exceed a certain score on one of these. Those who did not were denied diplomas, even if they met all other graduation requirements.

A group of 10 African-American students who were denied high school diplomas after failing three times to pass Florida's graduation test sued the state superintendent of education (Buckendahl and Hunt, 2007). The plaintiffs claimed that they had had neither adequate nor equal opportunity to master the "curriculum" on which the test was based. Ultimately, four different federal courtrooms would host various phases of the trial of *Debra P. v. Turlington* between 1979 and 1984.

"Debra P." won the case after a study revealed a wide disparity between what was taught in classrooms to meet state curricular standards and the curriculum embedded in the test questions. A federal court ordered the state to stop denying diplomas for at least four years while a new cohort of students worked its way through a revised curriculum at Florida high schools and faced a test aligned to that curriculum.

The *Debra P.* decision disallowed the use of NRTs for consequential, or "high-stakes", decisions. But, many states continued to use them for other purposes. Some were still paying for them anyway under multi-year contracts. Typically, states continued to use NRTs as systemwide diagnostic and monitoring assessments, with no consequences tied to the results.

Enter a young medical resident working in a high-poverty region of rural West Virginia in the mid-1980s. He heard local school officials claim that their children scored above the national average on standardized tests. Skeptical, he investigated further and ultimately discovered that every U.S. state administering NRTs claimed to score above the national average, a statistical impossibility. The phenomenon was tagged the "Lake Wobegon Effect" after Garrison Keillor's "News from Lake Wobegon" radio comedy sketch, in which "all the children are above average".

The West Virginia doctor, John Jacob Cannell, M.D., would move on to practice his profession in New Mexico and, later, California, but not before documenting his investigations in two self-published books, *How All Fifty States Are above the National Average* and *How Public Educators Cheat on Standardized Achievement Tests.* (Cannell, 1987, 1989)

---

[1] Such as the Iowa Tests of Basic Skills (ITBS), Iowa Test of Educational Development (ITED), Stanford Achievement Test (the "other SAT"), or the California Test of Basic Skills (CTBS).

Cannell listed all the states and all the tests involved in his research. Naturally, all the tests involved were nationally normed, off-the-shelf, commercial tests, the type that the *Debra P. v. Turlington* decision had disallowed for use with student stakes. It is only because they were nationally normed that comparisons could be made between their jurisdictions' average scores and national averages.

By the time Cannell conducted his investigation in the mid- to late-1980s, about twenty states had developed *Debra P*-compliant high-stakes state tests, along with state content standards to which they were aligned. But, with the single exception of a Texas test[2], none of them was comparable to any other, nor to any national benchmark. They were "criterion-referenced" or "standards-based" tests unique to each state, and not nationally norm-referenced tests. And, again with Texas excepted, Cannell did not analyze them.

Dr. Cannell cited educator dishonesty and lax security in test administrations as the primary culprits of the Lake Wobegon Effect, also known as "test score inflation" or "artificial test score gains".

With stakes no longer attached, security protocols for the NRTs were considered unnecessary, and relaxed. It was common for states and school districts to have purchased the NRTs "off the shelf" and handle all aspects of test administration themselves. Moreover, to reduce costs, they could reuse the same test forms (and test items) year after year. Even if some educators did not intentionally cheat, over time they became familiar with the test forms and items and could easily prepare their students for them. With test scores rising over time, administrators and elected officials discovered that they could claim credit for increasing learning.

Conceivably, one could argue that the boastful education administrators were "incentivized" to inflate their students' academic achievement. But, incentives exist both as sticks and carrots. Stakes are sticks. There were no stakes attached to these tests. In many cases, the administrators were not even obligated to publicize the scores. Certainly, they were not required to issue boastful press releases attributing the apparent student achievement increases to their own managerial prowess. The incentive in the Lake Wobegon Effect scandal was a carrot-specifically, self-aggrandizement on the part of education officials.

Regardless the fact that no stakes attached to Cannell's tests, however, prominent education researchers blamed "high stakes" for the test-score inflation he found (Koretz et al., 1991). Cannell had exhorted the nation to pay attention to a serious problem of educator dishonesty and lax test security, but education insiders co-opted his discovery and turned it to their own advantage (Phelps, 2006).

"There are many reasons for the Lake Wobegon Effect, most of which are less sinister than those emphasized by Cannell" (Linn, 2000, p.7) said the co-director of a federally-

[2] The Texas TEAMS was a hybrid, partly a complete NRT, but with other test items added to thoroughly cover state content standards. The NRT portion was used to make national comparisons. But, only items aligned to state content standards were used to make consequential decisions.

funded research center on educational testing-for over three decades the *only* federally-funded research center on educational testing.[3]

Another of the center's scholars added:

> *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.* (Koretz, 2008, p. 131)

These assertions supply 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, they present policymakers a conundrum: if scores on high-stakes tests improve, likely they are meaningless-leaving them no objective and reliable measure of school improvement. So they might just as well do nothing as bother doing anything.

After Dr. Cannell left the debate and went on to practice medicine, these education professors and their colleagues would repeat the mantra many times-high stakes (not lax security) cause test-score inflation-in dozens of reports published both by their center and by the National Research Council, whose educational testing research function they have co-opted (Baker, 2000; Linn, 2000; Linn, Graue, & Sanders, 1990; Shepard, 1990, 2000).

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 fingered "teaching to the test" for the crime, manifestly under pressure from the high stakes of the tests.

Cannell's tests had no stakes. That's a fact anyone can verify. The tests he included in his analysis are listed in his reports. Indeed, with the *Debra P.* decision settled in the federal courts in the early 1980s, Cannell's tests could not legally have had stakes. Nonetheless, ask most anyone inside education today for the primary lesson to emerge from Dr. Cannell's famous "Lake Wobegon Effect" studies, and they will tell you: high-stakes induces teaching to the test, which induces test-score inflation-artificial increases in test scores unrelated to actual gains in student learning.

On the one hand, it is astonishing that they stick with the notion because it is so obviously wrong. The SAT and ACT university admission tests have stakes-one's score on either helps determine which university one attends. But, they have shown no evidence of test-score inflation. (Indeed, the SAT was re-centered in the 1990s because of score *deflation.*) The most high-stakes tests of all-occupational licensure tests-show no evidence of test-score inflation. Both licensure tests and the SAT and ACT, however, have been administered with tight security and ample test form and item rotation.

---

[3] Since the early 1980s, the Center for Research on Educational Standards and Student Testing (CRESST) has been continually headquartered in UCLA's education school, and continually partnered with the University of Colorado's and the University of Pittsburgh's education schools. Other partners have included the Rand Corporation, and the education schools at Arizona State University, Stanford University, and at other University of California campuses.

3. Spot the Causal Factor

Table 1. Security and Stakes in evaluation

	HIGH SECURITY (EXTERNAL ADMINISTRATION)	LAX SECURITY (INTERNAL ADMINISTRATION)
High stakes	No test-score inflation e.g., SAT, ACT, licensure exams	Test-score inflation possible e.g., some internally administered district and state exams
No/low stakes	No test-score inflation e.g., National Assessment of Educational Progress (NAEP)	Test-score inflation possible e.g., Cannell's "Lake Wobegon" exams

Source: Auhor.

On the other hand, this "folk belief" is not unlike others in the US education school catechism, such as learning styles, multiple intelligences, and discovery learning: consistently proven wrong, but persisting nonetheless and matching the radical egalitarian and progressive education ideals that have consumed US schools of education.

The belief fits well into the knowledge base that many US education professors *want to* believe is true, rather than that which is true. US educationist doctrine may be less about a search for truth, and more an aspiration to what *should be* true -a set of knowledge they consider better because they consider it morally superior.

The late senator from New York, Daniel Patrick Moynihan, famously said "Everyone is entitled to their own opinion, but not their own set of facts"[4]. Apparently, US education professors do not agree. They have successfully elevated panoply of falsehoods aligned with their preferences to "facts" in the collective working memory. Their faux facts may influence US education policy-making more than real ones.

The scholars at the federally funded research center followed Cannell's studies with two of their own purporting to demonstrate both that teaching to the test works to artificially inflate test scores, and that high stakes induce teaching to the test. Both studies are methodologically flawed beyond the point of salvaging (Phelps, 2008, 2009a, 2010). Nevertheless, they remain, along with the distortion of Dr. Cannell's studies, highly respected among the US education professoriate and the foundation for most educators' understanding of the nature and implications of teaching-to-the-test (Crocker, 2005).

The reasoning goes like this: under pressure to raise test scores by any means possible, teachers reduce the amount of time devoted to regular instruction and, instead, focus on test preparation that can be subject-matter free (i.e., test preparation or test coaching). Test scores rise, but students learn less (Koretz, 1992, 1996; Koretz et al., 1991).

The two foundational studies examined certain patterns in the pre- and post-test scores from the first decade (i.e., late 1970s and early 1980s) of the federal government's compensatory education program (Linn, 2000) and the "preliminary findings" from the

[4] http://www.goodreads.com/author/quotes/219349.Daniel_Patrick_Moynihan

early 1990s of a test "perceived to be high stakes" in one school district (Koretz, Linn, Dunbar, & Shepard, 1991).

Research conducted on this hypothesis by others concludes that teachers who spend more than a brief amount of time focused on test preparation do their students more harm than good[5]. Their students score lower on the tests than do other students whose teachers eschew any test preparation beyond simple format familiarization and, instead, use the time for regular subject-matter instruction (see, for example, Allensworth, Correa, & Ponisciak, 2008; Camara, 2008; Crocker, 2005; Moore, 1991; Palmer, 2002). Moreover, students who know the specific content of prep tests beforehand may be lulled into a false confidence, study less, learn less, and score lower on final exams than those who do not (see, for example, Tuckman, 1994; Tuckman & Trimble, 1997).

The more widespread the belief that tests can be gamed by learning tricks unrelated to subject matter acquisition, the more customers and profits they gain.

As it turns out, neither of the two foundational studies of high-stakes testing effects included high-stakes tests. The researchers crossed their fingers behind their backs and employed an archaic, overly broad definition for the term "high stakes" for which virtually any standardized test would qualify (Phelps, 2010).[6] Yes, what they used was a definition, but it was neither the standard industry definition nor one that anyone outside their circle would reasonably assume for the term.[7]

This "floating definition" semantic sleight-of-hand is commonplace in US education research, its frequency of use grossly underappreciated by journalists and policy-makers. Education researchers surreptitiously substitute an obscure connotation for a term that varies from the more commonly understood denotation and explain the substitution, when they explain it at all, only in the fine print (Phelps, 2010).

One of the two studies was conducted in a school district and with tests that remain unidentified (Koretz, 2008). To this day, the researchers claim that they must keep that information secret to "protect" their sources (from what is not explained) (Staradamskis, 2008).

---

[5] Messick & Jungeblut (1981); DerSimonian & Laird (1983); Kulik, Bangert-Drowns, & Kulik (1984); Whitla (1988); Snedecor (1989); Becker (1990); Powers (1993); Allalouf & Ben-Shakhar (1998); Camara (1999); Powers & Rock (1999); Robb & Ercanbrack (1999); Briggs (2001); Zehr (2001); Briggs & Hansen (2004); Wainer (2011); Marte (2011); and Arendasy, Sommer, Gutierrez-Lobos, & Punter (2016).

[6] CRESST researchers cited (Shepard, 1990, p.17) a definition they attribute to James Popham from 1987 ascribing "high stakes" to any test whose aggregate results were reported publicly or which received media coverage. With the widespread passage of "truth in testing" and other open records laws, starting with California and New York State in the late 1970s, the aggregate results of all large-scale tests became public record. By their out-of-date definition, ALL large-scale tests are "high stakes".

[7] The standard, industry-wide definition of "high stakes" could be found in the *Standards for Educational and Psychological Testing* (AERA et al.), "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).

Secret definitions. Secret locations. Secret tests. Such studies may stand forever because they are neither replicable nor falsifiable. More like religion than science; they require faith. And, inside U.S. education one finds many willing believers.

Meanwhile, a cornucopia of studies contradicting the two research center studies have been repeatedly declared nonexistent by the same researchers and thousands of sympathetic others inside US education schools (Phelps, 2005, 2008, 2009b, 2012a, 2012b).

Elevating teaching-to-the-test to dogma, from the beginning with the distortion of Dr. Cannell's findings, has served to divert attention from scandals that should have threatened US educators' almost complete control of their own evaluation.[8] Had the scandal Dr. Cannell uncovered been portrayed honestly to the public-educators cheat on tests administered internally with lax security-the obvious solution would have been to externally manage all assessments (Oliphant, 2011).

More recent test cheating scandals in Atlanta, Washington, D.C., and elsewhere once again drew attention to a serious problem. But, instead of blaming lax security and internally managed test administration, most educators blamed the stakes and alleged undue pressure that allegedly ensues (Phelps, 2011a). Their recommendation, as usual: drop the stakes and reduce the amount of testing. Never mind the ironies: they want oversight lifted so they may operate with none, and they admit that they cannot be trusted to administer tests to our children properly, but we should trust them to educate our children properly if we leave them alone.

Perhaps the most profound factoids revealed by the more recent scandals were, first, that the cheating had continued for ten years in Atlanta before any responsible person attempted to stop it and, even then, it required authorities outside the education industry to report the situation honestly. Second, in both Atlanta and Washington, DC, education industry test security consultants repeatedly declared the systems free of wrongdoing (Phelps, 2011b).

Meanwhile, thirty years after J. J. Cannell first showed us how lax security leads to corrupted test scores, regardless the stakes, test security remains cavalierly loose in the United States. We have teachers administering state tests in their own classrooms to their own students, principals distributing and collecting test forms in their own schools. Security may be high outside the schoolhouse door, but inside, too much is left

[8] More than in most countries, the U.S. public education system is independent, self-contained, and self-renewing. Education professionals staffing school districts make the hiring, purchasing, and school catchment-area boundary-line decisions. School district boundaries often differ from those of other governmental jurisdictions, confusing the electorate. In many jurisdictions, school officials set the dates for votes on bond issues or school board elections, and can do so to their advantage. Those school officials are trained, and socialized, in graduate schools of education. A half-century ago, most faculties in graduate schools of education may have received their own professional training in core disciplines, such as Psychology, Sociology, or Business Management. Today, most education school faculty are themselves education school graduates, socialized in the prevailing culture. The dominant expertise in schools of education can maintain its dominance by hiring faculty who agree with it and denying tenure to those who stray. The dominant expertise in education journals can control education knowledge by accepting article submissions with agreeable results and rejecting those without. Even most testing and measurement PhD training programs now reside in education schools, inside the same cultural cocoon.

to chance. And, as it turns out, educators are as human as the rest of us; some of them cheat and not all of them manage to keep test materials secure, even when they aren't intentionally cheating.

4. Codifying TttT falsehoods

The primary advocates of the high-stakes-causes-TttT-which-causes-test-score-inflation belief (hereafter HS->TttT->TSI), reside at the Center for Research on Educational Standards and Student Testing (CRESST), for over three decades the only federally funded research center on educational testing. CRESST staff led the effort to discredit the work and findings of J. J. Cannell, the earnest medical doctor who uncovered the Lake Wobegon Effect scandal in the 1980s.

First, they rejected out of hand Cannell's contentions that educator cheating on tests was rampant and test security too lax. Second, in promoting HS->TttT->TSI, they instilled doubt in the reliability and validity of high-stakes test results.

Rather than stop there, however, they have advocated for thirty years that allegedly unreliable high-stakes test results should be "audited" by parallel low- or no-stakes tests. They reasoned that no-stakes test scores are reliable because there exist no incentives to cheat on them.

A cornucopia of research exists contradicting CRESST's faith in the reliability of low- and no-stakes test scores.[9] No matter, CRESST researchers have simply ignored it.

In summary, they promote all of the following beliefs:

1) HS->TttT->TSI. Again, as the theory's primary advocate writes,
2) "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".
3) Subject-matter independent training in test taking works to increase test scores (as some test prep companies also claim).
4) High-stakes test scores are, at best, only partly related to subject matter mastery, because they are also highly correlated with subject-matter-free test-taking skills.
5) The cause of educator cheating in testing administrations is high-stakes; without high-stakes, educators do not cheat.
6) No- or low-stakes tests, by contrast, are not susceptible to test-score inflation because there are no incentives to manipulate scores.

The public policy implications of these beliefs are substantial. Given the statements above, responsible public policy should incorporate the following:

---

[9] See, for example, Brown & Walberg (1993); Wise & DeMars (2005); Eklof (2007); Abdelfattah (2010); Barry, Horst, Finney, Brown, & Kopp (2010); Wise & DeMars (2010); Wainer (2011); Zilberberg, Anderson, Finney, & Marsh (2013); Steedle (2014); Liu, Rios, & Borden (2015); Sessoms & Finney (2015); Smith, Given, Julien, Ouellette, & DeLong (2015); Mathers, Finney, & Myers (2016); and Rios, Guo, Mao, & Liu (2016).

a) In the interest of improving test scores, teachers should teach to high-stakes tests—that is, drill on test format. They should reduce the amount of time devoted to subject matter mastery-to regular instruction and learning-and, instead, devote more time to taking practice tests, coaching students on test-taking strategies, familiarizing their students with standardized test formats, etc.

b) Use of test prep services should be encouraged. Moreover, in the interest of fairness, these services should be subsidized, at least for poorer students.

c) If score trends for high-stakes tests are unreliable and those for no- or low-stakes tests are reliable, no- or low-stakes tests may be used validly as shadow tests to audit the reliability of high-stakes tests' score trends.

d) Test security (or, the integrity of test materials) is not an issue with no- or low-stakes tests, so they can be validly administered without security controls.

e) Or, eliminate the use of high stakes tests entirely. Given that they provide neither valid nor reliable information, there is no excuse for using them. Currently, high stakes tests are used for certification and licensure in most professions and trades.

Several years ago, CRESST staff occupied prominent positions on the committee drafting an update to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), arguably the most important document in testing. Left in charge of drafting a new chapter 13 on public policy, CRESST staff incorporated their set of beliefs and the logical policy recommendations derived therefrom (Phelps, 2011b).

Judging from comments on the draft Standards made publicly available, I was the only public critic of the CRESST draft chapter 13. I recommended deleting it completely. As it turns out, an intervention occurred and the chapter was overhauled to remove the most egregious pseudoscientific claims and recommendations (AERA, NCME, & APA, 2013).

But, what if I hadn't raised a fuss? Did I represent the only barrier between the *Standards* incorporating CRESST's TttT Family of Fallacies and *Standards* based on genuine research evidence? That would be frightening. But, I witnessed no one else raising anything more than trivial objections to draft chapter 13.

A Pyrrhic victory? Meanwhile, the TttT Family of Fallacies has received warm welcomes at the Organisation for Economic Co-operation and Development (OECD) and the educational testing office at the World Bank (Phelps, 2014). These international organizations promote these falsehoods worldwide.

5. What if lax test security causes test score inflation?

Thousands of externally imposed high-stakes tests show no evidence of test-score inflation. Likewise, low- and no-stakes tests notoriously lead to test-score inflation when test security (or, "the integrity of test materials") is lax. The necessary and sufficient condition for test-score inflation is lax security, not high stakes.

Reject the pseudoscience of the TttT Family of Fallacies and quite different public policy implications emerge. Following where the research evidence points:

1) Test scores and test score trends should not be trusted in the absence of test security controls, no matter what the stakes.
2) High-stakes test scores and score trends are typically not only valid and reliable when administered with tight security, they are more likely to be valid and reliable because they are more likely to be administered with tight security than low- and no-stakes tests
3) Educators are normal human beings, and respond to a variety of incentives, just like the rest of us. By cheating on no- or low-stakes tests, educators might then publicize and take credit for the ostensible student learning increases. Note, however, that no "stakes" are involved; rather, self-aggrandizement is the motive.
4) Drilling on test format not only does not improve learning, because it takes time away from subject matter instruction, it reduces it.
5) Money spent on test preparation services is money wasted if the service consists primarily of test-taking strategies, format familiarity, and practice test taking.

Given the statements above, responsible public policy should incorporate the following:

a) Consider test security (or, the "integrity of test materials") far more seriously than it has been, and applicable to many no-or low-stakes tests.
b) Encourage teachers to devote only a modicum of time to familiarizing their students with standardized test-taking formats and strategies. They should not sacrifice instruction in subject-matter mastery.
c) Eliminate the fallacious research practice that considers no-stakes tests to be always valid and reliable and thus trustworthy to use in "auditing" high-stakes tests.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality, 38,* 159-168.

Allalouf A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*(1), 31-47.

Allensworth, E., Correa, M., & Ponisciak, S. (2008). *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 Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2013). *Standards for educational and psychological testing.* Washington, DC: AERA.

Arendasy, M. E., Sommer, M., Gutierrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence, 55,* 44-56.

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

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342-363. doi:10.1080/15305058.2010.508569

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

Briggs, D. C. (2001). The effect of admissions test preparation. *Chance*, *14*(1),10-18.

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

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research*, *86*(3), 133-136.

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* (pp. 147-158). Mahwah, NJ: Psychology Press.

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

Camara, W. J. (2008). College admission testing: Myths and realities in an age of admissions hype. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 45-76). Washington, DC: American Psychological Association.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools. How all fifty states are above the national average*. Daniels, WV: Friends for Education.

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

Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending standardized testing (*pp. 159-174). Mahwah, NJ: Psychology Press.

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

Eklof, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311-326. doi:10.1080/15305050701438074

Fraker, G. A. (1987). *The Princeton Review reviewed. The Newsletter*. Deerfield, MA: Deerfield Academy.

Gardner, W. (2008). *Good teachers teach to the test: That's because it's eminently sound pedagogy*. Retrieved from http://www.csmonitor.com/Commentary/Opinion/2008/0417/p09s02-coop.html

Koretz, D. (April, 1992). NAEP and the movement toward national testing. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Francisco.

Koretz, D. M. (1996). *Improving America's schools: The role of incentives*. Washington, DC: National Academy Press.

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

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (April, 1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

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

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (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.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79-94. doi: 10.1080/10627197.2015.1028618

Marte, J. (2011). *10 things test-prep services won't tell you. Market watch*. Retrieved from http://www.marketwatch.com/story/10-things-testprep-services-wont-tell-you-1301943701454

Mathers, C., Finney, S., & Myers, A. (2016, July). *How test instructions impact motivation and anxiety in low-stakes settings*. Paper presented at the Annual Meeting of the Psychometric Society, Asheville, NC.

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

Moore, W. P. (1991). *Relationships among teacher test performance pressures, perceived testing benefits, test preparation strategies, and student test performance* (PhD dissertation). University of Kansas, Lawrence.

Oliphant, R. (2011). Modern metrology and the revision of our Standards for Educational and Psychological Testing: An open letter to American parents. *Nonpartisan Education Review / Essays*, *7*(4). Retrieved from http://www.nonpartisaneducation.org/Review/Essays/v7n4.pdf

Palmer, J. S. (2002). *Performance incentives, teachers, and students: Estimating the effects of rewards policies on classroom practices and student performance* (PhD dissertation). Ohio State University, Columbus, Ohio.

Phelps, R. P. (2005). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55-90). Mahwah, NJ: Psychology Press.

Phelps, R. P. (2006). A tribute to John J. Cannell, M.D. *Nonpartisan Education Review/Essays*, *2*(4). Retrieved from http://www.nonpartisaneducation.org/Review/Essays/v2n4.pdf

Phelps, R. P. (2008/2009a). The rocky score-line of Lake Wobegon. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp.102-134). Washington D. C.: American Psychological Association.

Phelps, R. P. (2008/2009b). Educational achievement testing: Critiques and rebuttals. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 66-90). Washington D. C.: American Psychological Association.

Phelps, R. P. (2010). The source of Lake Wobegon. *Nonpartisan Education Review/Articles, 6(3)*. Retrieved from http://nonpartisaneducation.org/Review/Articles/v6n3.htm

Phelps, R. P. (2011a). *Standards for Educational & Psychological Testing*. New Orleans, LA: American Psychological Association.

Phelps, R. P. (2011b). Educator cheating is nothing new; doing something about it would be. *Nonpartisan Education Review/Essays, 7*(5). Retrieved from http://nonpartisaneducation.org/Review/Essays/v7n5.htm

Phelps, R. P. (2011c). *Teach to the test? The Wilson Quarterly*. Retrieved from http://wilsonquarterly.com/quarterly/fall-2013-americas-schools-4-big-questions/teach-to-the-test/

Phelps, R. P. (2012a). Dismissive reviews: Academe's Memory Hole. *Academic Questions, 25*(2), 228-241.

Phelps, R. P. (2012b). The rot festers: Another National Research Council report on testing. *New Educational Foundations, 1*(1). Retrieved from http://www.newfoundations.com/NEFpubs/NewEduFdnsv1n1Announce.html

Phelps, R. P. (2014). Synergies for better learning: An international perspective on evaluation and assessment. *Assessment in Education: Principles, Policies, & Practices, 21*(4), 481-493. doi:10.1080/0969594X.2014.921091

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

Popham, W. J. (2004). All about accountability / "Teaching to the test": An expression to eliminate. *Educational Leadership, 62*(3), 82-83.

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

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

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 16*, 1-36. doi:10.1080/15305058.2016.1231193

Robb, T. N., & Ercanbrack, J. (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, 3*(4).

Sessoms, J., & Finney, S. J. (2015) Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*(4), 356-388. doi:10.1080/15305058.2015.1034866

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15-22.

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

Smith, J. K., Given, L. M., Julien, H., Ouellette, D., & DeLong, K. (2013). Information literacy proficiency: Assessing the gap in high school students' readiness for undergraduate academic work. *Library & Information Science Research*, *35*, 88-96.

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.

Staradamskis, P. (2008, Fall). Measuring up: What educational testing really tells us. Book review. *Educational Horizons*, *87*(1). Retrieved from http://nonpartisaneducation.org/Foundation/KoretzReview.htm

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education*, *27*, 58-76. doi: 10.1080/08957347.2013.853072

Tuckman, B. W. (April, 1994). 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.

Tuckman, B. W., & Trimble, S. (August, 1997). Using tests as a performance incentive to motivate eighth-graders to study. Paper presented at the *Annual Meeting of the American Psychological Association*, Chicago.

Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies.* Princeton, NJ: Princeton University Press.

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

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1-17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41. doi:10.1080/10627191003673216

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

Zilberberg, A., Anderson, R. D., Finney, S. J., & Marsh, K. R. (2013). American college students' attitudes toward institutional accountability testing: Developing measures. *Educational Assessment*, *18*, 208-234. doi:10.1080/10627197.2013.817153

Breve CV del autor

Richard P. Phelps

Founder of the Nonpartisan Education Group and editor of its peer-reviewed journal, the *Nonpartisan Education Review* (http://nonpartisaneducation.org), a Fulbright Scholar, and a fellow of the Psychophysics Laboratory. He has authored, or edited and authored, four books on assessment policy –*Correcting Fallacies about Educational and Psychological Testing* (APA); *Standardized Testing Primer* (Peter Lang); *Defending Standardized Testing* (Psychology Press); and *Kill the Messenger: The War on Standardized Testing* (Transaction)– and several statistical compendia. Phelps has held positions with several organizations working in assessment, including ACT, AIR, ETS, the OECD, Pearson, and Westat. He holds degrees from Washington, Indiana, and Harvard Universities, and a PhD in Public Policy from the University of Pennsylvania's Wharton School. ORCID ID: 0000-0003-4008-087x. Email: richardpphelps@yahoo.com