Citation: Phelps, R.P. (2016, January). Teaching to the test: A very large red herring. *Nonpartisan Education Review/Essays, 12*(1).

Teaching to the test: A very large red herring[1]

Richard P. Phelps[2]

Standardized testing is one of the few means by which the public may ascertain what transpires inside our country's 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).

As criticism, teaching to the test suggests 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?

What if we all agreed that teaching to the test was bad practice, where would that leave the teacher? Should teachers purposely not teach material that will be tested? If a test is aligned with those standards, and its questions thoroughly cover them, can responsible teachers avoid teaching to the test? (Gardner 2008)

But, the meaning of the phrase is slippery (Shepard 1990, p. 17; 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".

A more viable teaching-to-the-test criticism concerns teaching in a manner that is not considered optimal for learning standard content or skills, but is believed to improve test performance. Instruction on standardized test formats: such as multiple-choice, drilling with test-maker-provided workbooks, or administering practice tests are examples (Shepard 1990, p. 19).

But, teaching to the test is far more than a catch phrase or slogan. It has served for three decades to divert attention from an endemic problem—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 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)[3] (Phelps 2008/2009; 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, 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)

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[4], none of them was comparable to any other, nor to any national benchmark. 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 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. But, there were no stakes attached to these tests. In many cases, the administrators were not 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, p.2). 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-funded research center on educational testing—for over three decades the *only* federally-funded research center on educational testing.[5]

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 (Linn, Graue, & Sanders 1990; Shepard 1990; Baker 2000, p.18; Linn 2000, pp. 5, 7; Shepard 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 have stakes—one's score on either helps determine which college 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.

Spot the Causal Factor

	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

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 US education professors *want to* believe is true, rather than that which is true.

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."[6] Apparently, US education professors do not agree. They have successfully elevated a 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 induces 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, pp. 2, 3).

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, 5, 6) and the "preliminary findings" from the 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[7]. Their students score lower on the tests than do other students whose teachers eschew any test preparation beyond simple format familiarization and, instead, use the time for regular subject-matter instruction (see, for example, Moore, 1991; Palmer, 2002; Crocker, 2005; Camara, 2008; Allensworth, Correa, & Ponisciak, 2008). 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).

Proponents of the high-stakes -> teaching to the test -> score inflation belief, however, have allies among the coterie of private sector test prep companies (Fraker 1986-87; Smyth 1990). 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).[8] 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.[9]

This "floating definition" semantic sleight-of-hand is commonplace in 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).

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 education schools (Phelps 2003, 2005; 2008/2009; 2012a; 2012b).

Education scholars who manage to establish an appealing falsehood as fact in public belief systems rank among the most highly rewarded in the profession. The primary salesperson for <high-stakes -> teaching to the test -> score inflation> retains an endowed chair at Harvard University. It is in the education school, but it is still Harvard. In History, Physics, or Mathematics, Harvard and Stanford may, indeed, host the country's most deserving scholars. In Education, they host some of country's cleverest obfuscators.

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.[10] 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).

Recent test cheating scandals in Atlanta, Washington, DC, 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 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. 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 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.

Citation: Phelps, R.P. (2016). Teaching to the test: A very large red herring. Nonpartisan Education Review/Essays, 12(1). http://nonpartisaneducation.org/Review/Essays/v12n1.pdf

Endnotes

[1] According to Literary Devices "Red herring is a kind of fallacy that is an irrelevant topic introduced in an argument to divert the attention of listeners or readers from the original issue. In literature, this fallacy is often used in detective or suspense novels to mislead readers or characters or to induce them to make false conclusions."

[2] Copyright 2016, Richard P. Phelps.

[3] 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)

[4] 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.

[5] 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.

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

[7] 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; and Arendasy, Sommer, Gutierrez-Lobos, & Punter, 2016.

[8] 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".

[9] 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)

[10] 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.

References

Allalouf A., & Ben-Shakhar, G. (1998, March). 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, 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.

Arendasy, M.E., Sommer, M., Gutierrez-Lobos, K., & Punter, J.F. (2016, March-April). 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*. William H. Angoff Memorial Lecture. Educational Testing Service, Policy Information Center, Princeton, NJ.

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

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.

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.

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.

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.

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

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

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

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.

McCall, W. A. (1939). *Measurement*. New York: The Macmillan Company.

Messick, S. & A. Jungeblut (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 [date] 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. Columbus, Ohio: Ohio State University.

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

Phelps, R. P. (2008/2009a). The rocky score-line of Lake Wobegon. Appendix C in R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association. http://supp.apa.org/books/Correcting-Fallacies/appendix-c.pdf

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

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

Phelps, R.P. (2011a, April 10). Extended Comments on the Draft *Standards for Educational & Psychological Testing* (But, in particular, Draft Chapters 9, 12, & 13) to the Management Committee, American Psychological Association, National Council on Measurement in Education, and American Educational Research Association, New Orleans, LA

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

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

Phelps, R. P. (2012a, June). Dismissive reviews: Academe's Memory Hole. *Academic Questions*, *25*(2), pp. 228–241. http://www.nas.org/articles/dismissive_reviews_academes_memory_hole

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

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

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

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). http://nonpartisaneducation.org/Foundation/KoretzReview.htm

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. Chapter II in G.M. Whipple (Ed.), *The Seventeenth yearbook of the National Society for Study*

of Education. Part II. The Measurement of Educational Products. Bloomington, IL: Public School Publishing Co.

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).

Wainer, H. (2011, pp.134-137). *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.

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