Dan Koretz's Big Con

The Testing Charade: Pretending to Make Schools Better, by Daniel Koretz [book review]

Reviewed by Richard P. Phelps

The mainstream research that informs our country's education policies is often more caricature than genuine research. Policy discussions tend to be dominated by the research that the ruling class inside education *wishes to be* true, rather than by that which is true.

Among the several falsehoods book author Daniel Koretz and his colleagues have peddled over the years is the claim that the evidence for the benefits of testing is "thin" (and the evidence for costs abundant). Largely in response to their claims, I several years ago published a meta-analysis of 100 years' worth of research on the effects of testing on student achievement. I reviewed over 800 quantitative, experimental, survey, and qualitative studies. The weight of the latter two types of studies was overwhelmingly positive (e.g., 93% of qualitative studies found a positive result to a testing intervention and the average effect size for survey studies exceeded 1.0, a very high effect. The effect sizes for the quantitative and experimental studies—hundreds of mostly random assignment experiments dating back to the 1920s—ranged between moderately and highly positive.[1]

Because I read and heard the same messages as everyone else from those prominent education researchers who receive press attention, I had expected to find clearly negative effects. Some of the most widely covered studies allegedly demonstrating that testing was, on balance, harmful were included in my meta-analysis. But also included in my meta-analysis were hundreds of studies that had received virtually no public attention. Testing experts, education practitioners, and psychologists performed most of those studies.

(True to form, not a single education journalist has ever asked me about the meta-analysis. Meanwhile, DC-based education journalists talk to anti-testing spokespersons thousands of times a year and often promote the single research studies conducted by celebrity researchers as hugely consequential to policy.)

Therein lies the chief secret of the success of the anti-testing forces in education research: they count (i.e., cite or reference) the research that reaches anti-testing conclusions and they ignore the abundance of research that contradicts. (For the few pro-testing studies that receive so much public attention they cannot simply ignore them, other information suppression methods may be used, such as dismissive reviews, tone policing, misrepresentation, or character assassination).

Harvard Education Professor Daniel Koretz is no different. If one chooses to assume that the tiny sample of the research literature on educational testing conducted by him and a small coterie of sympathetic colleagues represents the universe of the relevant research literature, his new book, *The Testing Charade: Pretending to Make Schools Better*, makes some sense. Consider the entire breadth of the research literature on testing, however, and it makes no sense at all.

Ironically, in regard to tests, Koretz declares,

> "There is no way to test the entire domain. There just isn't time, even with the excessive amount of time many American schools now devote to testing. So we test a small part of the domain and use the tested part to estimate how well students would have done if we had tested the whole thing."
>
> "…the sampling of content needed to create a test--is the single most important thing to understand about standardized testing, and it is the root of many problems that I describe in the following chapters."

In truth, most standard-based tests test their entire domain. Unlike Koretz, I have worked in test development for organizations that created standards-based tests. It was imperative with each test we developed that each and every standard was tested with at least one test item. If a standard suggested that students know how to add two digit numbers, we included at least one test item that required the addition of two-digit numbers, and so on down the complete list of all the standards that teachers had covered.

In a public presentation of his new book, Koretz gave the example of a 12th grade math test containing only the typical number of items. Koretz declared that this quantity of items was used to measure 12 years of mathematics education.[2]

That claim, too, is disingenuous. The 12th grade test was, more likely, designed to measure mastery of the 12th grade math standards. It didn't

need to measure mastery of the standards for each of the other grade levels. The year-end tests in those grade levels had already done that. To make his claim about domain coverage valid, one needs to imagine as credible a 12th grade test that includes first and second grade math test items.

At the same presentation, Koretz declared that the stakes of tests were "chicken feed" back in the late 1980s compared to those today. Again, he is exactly wrong. In the late 1980s, over twenty states administered high school graduation exams, for which failure meant a diploma withheld. Today's federally required tests portend no consequences for students whatsoever. And, some of the states that have continued to administer high school exit exams are now abandoning them.

Koretz's primary claim to fame, of course, is his test score inflation hypothesis, which he claims is proven by decades of research. Indeed, test score inflation is real; it is an artificial increase in test scores over time that is unrelated to an increase in learning. And, it is common.

The problem of test score inflation was first brought to light by John J. Cannell's "Lake Wobegon Effect" studies. Doctor Cannell found every U.S. state that administered national norm-referenced tests claiming to be "above average," a statistical impossibility.

For years, Koretz and his colleagues cited Cannell's studies as evidence for the high-stakes-causes-test-score-inflation claim. Indeed, at the time—the mid to late 1980s—dozens of states administered high-stakes tests. With Koretz and his colleagues at the Center for Research on Education Standards and Student Testing (CRESST) promoting the idea that high stakes caused the score inflation that Cannell found, most in U.S. education came to believe it.

In his new book, Koretz doesn't even mention Cannell's work, instead claiming that he himself conducted the first-ever empirical study of score inflation, in the years after Cannell's. Koretz also now admits that the alleged "high stakes" test in his own famous late 1980s study was not high stakes at all. It was a no-stakes test that he has decided to call high stakes, contrary to the clear definition of the term. (Perhaps my protests over the past decade at Koretz's misrepresentation of the historical record have had some effect.)

Cannell didn't study score trends for all the tests administered by states in the mid- to late-1980s, though. He studied only those that were nationally normed. And, with the exception of only a single state, none of his score-

inflated tests had any stakes. They were no stakes diagnostic or monitoring tests.

Koretz has consistently argued that no-stakes test score trends are consistent over time ("because there is no incentive to manipulate scores"). So, consistent, in fact, that one can use no-stakes test score trends to shadow the trends of the allegedly more manipulable high-stakes tests.

The actual results of Cannell's studies, then, present embarrassing evidence for Koretz—directly contradicting his primary contribution to the education research literature. Cannell found publicly announced test scores on no-stakes tests rising over time such that all states ended up above the original national average.

Given stakes were not involved, how did the scores rise over time? Cannell fingers cheating and, admittedly, there was plenty of that. But, more precisely, the cause was lax test security. Unlike high-stakes tests, which are more likely to be administered with tight security, no-stakes tests are typically administered with no test security. Schools were reusing the same test forms year after year. Teachers were studying the test forms prior to test administration.

So, of course, there was test-score inflation. Teachers and students could anticipate the actual content of the tests. Add the facts that educators themselves determined the test administration procedures, could score the tests themselves, and completely controlled how the test scores were reported to the public. The same upper-level administrators then used the inflated scores as evidence of their own managerial prowess. The incentive for test-score inflation was self-aggrandizement, not the alleged perverse incentives of high stakes.

If one searches, one will find that every high-stakes test Koretz has identified as score-inflated was also administered with lax security. Meanwhile, there exist thousands of high-stakes tests administered with tight security that show no evidence of score inflation. And, as Cannell showed us, there exist plenty of score-inflated no-stakes tests.

Koretz recommends that "auditing" of allegedly unreliable high stakes tests be done with parallel no-stakes tests administered over the same time period. But, the no-stakes test used most of the time for this purpose is the National Assessment of Educational Progress (NAEP). The NAEP is indeed a no-stakes test; it is also one of the few no-stakes tests administered externally with tight security.

Tests are administered "externally" when organizations unaffiliated with the schools administer them. The adults proctoring NAEP tests work for NAEP, not the schools where the tests are administered.

Currently (November 26, 2017), I am in Chile, where I am consulting on issues related to the nationwide university admission test. Tomorrow, that test will be administered countrywide by 25,000 proctors, all of whom have been trained and will be paid by the agency that develops and administers the test. School personnel never see the tests, they are not available prior to administration, and they are completely different from year to year. As is true with all tests administered externally with tight security and ample test form and item rotation, Chile's test has never shown evidence of score inflation.

Meanwhile, in the United States, test security is about as lax as can be. The scandal that Cannell's study should have caused was successfully muted by the dishonest spin provided by Koretz and his CRESST colleagues. Why might U.S. educators oppose secure externally administered high stakes tests? Because they do not control them; and they might be used to judge their performance.

Koretz aids educators' intransigence with the following themes:

- if high-stakes test scores can not be trusted, policymakers might as well not test, given they do more harm than good;
- if no-stakes test scores are trustworthy and reliable, they also happen to be "internally" administered (i.e., under the control of educators to freely manipulate);
- if there exists no evidence of high-stakes testing benefits, there is no reason to look for it

Daniel Koretz is an intelligent human being. He must realize that he is misleading the public. The fact that he has gotten away with the deception for a quarter century indicates the lowly depth to which US education research has sunk. He hasn't pulled it off by himself, though. There exist plenty of education researchers who know that Koretz's claims are bunk, and I sometimes hear from them in private. But, those willing to speak out publicly, and risk their careers by opposing popular education establishment dogma, remain few and far between.

Note: The Appendix below contains point-by-point rebuttals of dozens of Koretz's statements.

Citation: Phelps, R.P. (2017). Dan Koretz's Big Con, *Nonpartisan Education Review* / Reviews. Retrieved [date] from http://nonpartisaneducation.org/Review/Reviews/v13n1.htm

---

For more on this topic:

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. http://nonpartisaneducation.org/Review/Books/CannellBook1.htm

Cannell, J.J. (1989). *How Public Educators Cheat on Standardized Achievement Tests: The "Lake Wobegon" Report*. Albuquerque, NM: Friends for Education. http://nonpartisaneducation.org/Review/Books/Cannell2.pdf

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

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

Phelps, R. P. (2008/2009). 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. (2008/2009). 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. (2011, 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. http://nonpartisaneducation.org/Review/Essays/v7n3.htm

Phelps, R. P. (2011, 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. (2011, 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. (2012, 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. (2012, October 2). The School-Test Publisher Complex. *Education News*. http://www.educationnews.org/education-policy-and-politics/richard-p-phelps-the-school-test-publisher-complex/

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

Phelps, R.P. (2014). The gauntlet: Think tanks and federally funded centers misrepresent and suppress other research. *Nonpartisan Education Review/Essays, 10(1).* http://nonpartisaneducation.org/Review/Essays/v10n1.htm

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

Phelps, R.P. (2017). "Teaching to the test" family of fallacies. *Revista Iberoamericana de Evaluación Educativa, 10*(1). https://revistas.uam.es/index.php/riee/article/view/7593

Phelps, R.P. (2017). It's a myth: High stakes cause test score inflation. Paper presented at the ResearchED-US Conference, Brooklyn, NY, October 7, 2017. https://www.slideshare.net/richard_p_phelps/its-a-myth-high-stakes-cause-test-score-inflation-80623917?qid=ec66953a-c77b-4ab7-8726-9d31b2da97af&v=&b=&from_search=1

Appendix

In the table below, I respond to some of Koretz's claims point by point. As I read the Kindle version, I provide "location" rather than page numbers for the quotes from his book.

Kindle location	Koretz quote	Response
65	"Newspapers are replete with reports of students who are so stressed by testing that they become ill during testing or refuse to come to school. … many children cried … others vomited or lost control of their bowels or bladders. Others simply gave up. One teacher reported that a student kept banging his head on the desk, and wrote, 'This is too hard' ...throughout the test booklet."	American students are some of the least academically challenged in the world. If they are stressed, it is not likely to be from academic pressure.
142	"However, this reasoning isn't just simple, it's simplistic--and the evidence is overwhelming that this approach [that testing can improve education] has failed. … these improvements are few and small. Hard evidence is limited, a consequence of our failure as a nation to evaluate these programs appropriately before imposing them on all children."	In truth, the evidence is overwhelming that testing can improve education. Koretz just won't tell you about it. He is highly selective of the research evidence he provides and has no qualms about declaring nonexistent the thousands of studies that contradict his claims.
152	"This problem was predicted by measurement experts nearly seventy years ago, and we have more than twenty years of research showing that false gains are common and often very large."	"Twenty years of research?" Some of the studies he cites administered no-, not high-stakes, tests. All th studies involved lax test security, for which no experimental controls were implemented.
225	"There is no way to test the entire domain. There just isn't time, even with the excessive amount of time many American schools now devote to testing. So we test a small part of the domain and use the tested part to estimate how well students would have done if we had tested the whole thing."	"Entire domains" are tested all the time. Unlike Koretz, I worked in test development. For standards based tests--what most state tests are--we wrote at least one test item for each and every standard. ALL the standards were tested, and they comprised the entire domain.
225	"Most of the domain remains untested, just as most voters are not reached by pollsters."	It is a false analogy. Yes, most polls sample. Most standards-based tests, by contrast, do not. They are censuses, not samples, of the associated content standards.

258	"Her point was that she had noticed that the item-writers for the state test always happened to use regular polygons."	Good item writers are not so predictable. The solution to this perceived problem is to hire competent professionals to write and evaluate test items, and administer tests with some variety of item types.
258	"What I have just explained--the sampling of content needed to create a test--is the single most important thing to understand about standardized testing, and it is the root of many problems that I describe in the following chapters."	And, for standards-based tests--the type that most high-stakes tests are--it's a falsehood.
269	"One of the worst in this respect is the 'value-added' estimates used to evaluate teachers in many states. These are highly imprecise, …"	Koretz should admit responsibility for his part in the value-added measurement (VAM) boondoggle. VAM relies on the assumption that one can validly compare student performance on tests with no stakes for them. Koretz is the world's leading advocate of the notion that no-stakes test scores ar reliable and comparable, "because there is no incentive to manipulate the scores."
301	"The bottom line: the information yielded by tests, while very useful, is *never* by itself adequate for evaluating programs, schools, or educators. Self-evident as this should be, it has been widely ignored in recent years. Indeed, ignoring this obvious warning has been the bedrock of test-based education reform."	I know of no testing professional who claims that testing by itself is adequate for evaluating programs schools, or educators. But, by the same notion, neither are other measures used alone, such as inspections or graduation rates.
324	"This is the core of inappropriate test preparation, which is epidemic in American schools today."	And, he should look in the mirror when he tries to pass blame for this "epidemic." Many have trusted him when he says that inappropriate test preparation garners higher test scores. The evidenc does not support him.
324	"As long as the details of sampling are predictable--and test prep companies are very good at finding predictable patterns in tests--teaching to the specifics of the test will inflate scores."	Again, most standards-based tests do not sample, they test each and every standard, and fairly unambiguously. Much commercial test prep is frauc If tests are administered externally, with tight security and ample form and item rotation, educators cannot teach to the specifics of the test, because they do not know them.

343	"This test-centered world has been in place long enough, in various forms, that many people think it is the normal state of affairs. It isn't. It is relatively new in the United States."	Koretz has a point here. But, he advocates that we return to the days of the Lake Wobegon Effect, whe all tests were "internally" administered and educators were free to manipulate any and all aspects of their administration and reporting. Koret wishes to return to the practices that John J. Cannel exposed.
366	"Yet even these countries, in which high-stakes testing exerts more pressure on students than it does here, don't do what we do. The don't test students nearly as frequently as we do, and they don't give test scores such a large and direct role in evaluating schools and teachers."	The key word here is "frequently." It has become common in the US to administer many short tests with no consequences for students. Other countries are more test-efficient. They may use fewer tests overall, but those tests count and they can be very long--several days or weeks.
386	"But Finland, in recent years the darling of the education world because of its very high performance on one of the international assessments in math, and Germany have no high-stakes testing in any other grades."	In truth, both Finland and Germany have high-stake testing in other grades.
415	"In addition to expanding the amount of testing and increasing its importance, the reforms of the 1980s brought another important change from the minimum-competency era: the shifted the focus away from holding student accountable for their scores to using students' scores to hold educators directly accountable."	If by "educators" he means to imply teachers, this is wrong. There were zero states using test scores or test score trends to evaluate teachers in the 1980s. …or the 1990s. Tennessee, the state that pioneered value-added measurement, did calculate teacher ratings based on student test score trends, but ther were no negative consequences for teachers at any time in the 20th century.
529	"…other subjects, such as history, civics … aspects of math and reading that are hard to measure with standardized test …"	It is not at all difficult to measure history or civics knowledge with standardized tests. Have no idea what aspects of math or reading he refers to here a not testable.
550	"The Dutch do use standardized testing to evaluate schools, but they test far less than we, …:	The Dutch do not test far less than we do; in fact they test more than we do--every year, starting in grade 1 and in more subjects--not just reading and math like here--and at more grade levels. You can even consult the source Koretz cites to see how wrong he is.
560	"Absolutely no one is given any incentive to monitor or control how these gains are achieved."	…so long as tests (and test security protocols) are administered by school personnel.

563	"It is no accident that many of the types of inappropriate test prep by teachers that I will describe in coming chapters were actively encouraged by administrators, and it is no accident that some of the most publicized cheating scandals were actively supported by--and in at least one case deliberately supported by--the people at the top."	To eliminate most inappropriate test prep, deny school personnel any access to test materials before actual administration. To eliminate the practice of school personnel altering answer sheets, deny them access to answer sheets. Some cheating prevention measures are remarkably simple, such as not allowing teachers to proctor tests with their own students or posting two proctors in each testing room. Even these simple, obvious measures are not implemented because schools are not told to implement them.
723	"Although this problem has been documented for more than a quarter of a century, it is still widely ignored, and the public is fed a steady diet of seriously misleading information about improvements in schools."	The public is fed a steady diet of misleading information from their schools because schools control the administration of the tests used to evaluate them. This will never change until evaluative tests are administered by independent third parties (e.g., the state auditor's office).
784	"I told them that they were right to worry about possible inflation but that I was aware of no malfeasance and would have been greatly surprised if there was any."	He shouldn't have been surprised if he genuinely understood how the New York tests were administered and scored, with lax security.
831	"Don't think this was only a New York problem. Superintendents and commissioners generally aren't eager to have studies of possible score inflation in their systems. Trust me: asking one of them for access to their data in order to find out whether scores are inflated doesn't usually get a welcoming response. So there are far fewer audits of impressive score gains on high-stakes tests than there ought to be."	No, we shouldn't trust him. Externally administered high-stakes testing is widely reviled among US educationists. It strains credulity that Koretz cannot find one district out of the many thousands to cooperate with him to discredit testing.
841	"…as of the late 1980s there was not a single study evaluating whether inflation occurred or how severe it was. With three colleagues, I set out to conduct one."	Wrong. John J. Cannell's mid 1980s "Lake Wobegon Effect" study showed all US states claiming to be "above average" on nationally normed tests. Perhaps Koretz doesn't mention Cannell here because all but one of Cannell's score-inflated tests had no stakes. They were no-stakes tests administered with minimal security. The cause of test score inflation is lax security, not high stakes. It was Cannell's honest and objective study that prompted Koretz to do his.

852	"We designed the experiment … wrote some tests ourselves, purchased materials for two commercial tests."	It was hardly an "experiment." There were no controls. The "two commercial tests" phrase is informative, though. The ultimate test score comparison was made between norm-referenced tests from two different test publishers. All possible such comparisons were shown at the time to be invalid due to poor alignment between competing tests.
863	"In addition, we were not even allowed to identify the specific tests used because that might provide a clue about the district's identity."	It is difficult to believe such secrecy is still needed, i it ever was, a quarter century later.
894	"This second test, usually called an audit test, should be a credible test that is unaffected … by score inflation. NAEP is the most commonly audit test for several reasons. It is considered a very high-quality test. NAEP scores are not susceptible to inflation because teachers aren't held accountable for scores and therefore have no incentive to engage in NAEP-focused test prep."	Of more significance: The NAEP is externally administered (i.e., NOT by school personnel), under tight security. It is one of very few no-stakes tests that is.
926	"…to a different but very similar test (Test B),…"	"Very similar"? Not hardly. It was a "competing" national norm-referenced test, developed by a different company using completely different assumptions about course topics, topical emphases and the sequencing of topics across the grades. Other researchers compared the content of these "competing" tests and found little alignment, declaring the tests non comparable. (E.g., the Iowa Test for grade 4 math did not align with the California Test for grade 4 math.)
939	"That's why we present only the third-grade results. The fifth-grade results were more extreme, but they were potentially misleading. (The reason is complex, but motivation isn't a serious threat to the design of most other studies of score inflation."	Motivation is a mortal threat to the type of score inflation study Koretz tries to do. Different students exert varying effort on no-stakes tests. Effort varies by age, gender, ethnicity, and so on. No-stakes test scores are less reliable or comparable than high-stakes test scores.
961	"However, all that is required for scores to become inflated is that the sampling used to create a test has to be *predictable*."	Again, with most standards-based tests, there is no sampling. All the standards are tested. If there is sampling employed for another type of test, that sampling should not be predictable.

972	"While the system in the-district-I-cannot-name was high stakes by the standards of the time, it was extremely lenient compared with what we now impose on schools."	Nonsense. Stakes were higher for students at the time of Koretz's study. Over 20 states administered high-stakes high school exit exams. Some states administered high-stakes tests in other grades, too. Today. With more and more states dropping their high school exit exams, many students now face ZERO high-stakes tests. For students, testing stakes have declined and continue to decline further.
972	"For the most part, scores did not lead to concrete consequences for either students or teachers. Students were not held back in grade because of low scores. Teachers and principals, at least if tenured, were not at risk of losing their jobs because of scores, and tenure was not based on scores. There were no cash rewards...."	Thus, Koretz's late-1980s study has been misrepresented for a quarter century. By any definition anyone currently uses, Koretz's key test was not high stakes.
1003	"However, value-added estimates are rarely calculated with lower-stakes tests that are less likely to be inflated."	Nonsense. All value-added estimates are calculated with tests that have no stakes for the students. Moreover, lower-stakes tests are more likely to be inflated than higher-stakes tests.
1023	"They know that test prep can sometimes produce much faster gains, but they don't care."	The preponderance of evidence does not support Koretz's assertion that test prep produces much faster gains. That may be why he hedges his statement with "sometimes." Well, just about anything can happen "sometimes."
1045	"Some of you may remember the 'Texas miracle' of the 1990s: scores went up very rapidly on the state test … were illusory, NAEP ... far smaller gains overall…"	The Texas gains were not illusory; between 1990 an 1998 Texas' NAEP scores improved more than any other state's, bar one. The study by Koretz's colleagues, however, was misrepresented by its ow authors. They performed hypothesis tests separatel on each NAEP test (in each subject, at each grade level), then made statements about Texas educatio as a whole. To make valid conclusions about the state as a whole they should have pooled the result from all grades and subjects. Had they done so, the results would have been highly statistically significant.

1122	"In 2006, a young math teacher at Parks Middle School in Atlanta, Damany Lewis, went into a locked room where state tests were stored and removed a copy in order to provide teachers in the school with advance knowledge of the questions."	So, Koretz would argue, eliminate high-stakes testing. Most anyone outside the education industr would say, administer the tests externally--i.e., not with school personnel--and tighten security. As is done with the NAEP.
1264	"…how difficult it can be to verify cheating and how resistant the system can be to thorough and honest investigation. There are several reasons why cheating is not investigated carefully more often than it is, including the financial cost and a tendency to assume that no one is misbehaving unless the data shout that someone is."	Koretz neglects to mention that in most of the United States, the cheaters are the same people responsible for investigating cheating. What is needed is external administration of tests, and external investigations.
1424	"It is worth considering why we are so unlikely to ever find out how common cheating has become. … the press remains gullible…"	The press remains gullible to Koretz's myth making, too.
1648	"The problem with POE [process of elimination] is that some of the student who find the correct answer by eliminating incorrect ones would be unable to generate the correct answer if they weren't given alternatives from which to select. … And once they leave school and enter the read world, that's what they will usually encounter."	Multiple-choice and constructed-response items each have advantages and disadvantages, which is why most testing experts suggest using both in a te when possible. But, Koretz is wrong about process o elimination (POE). We use it in our daily lives all the time. One underestimated advantage of multiple-choice items in tests is their accumulation of partial credit for partial knowledge. If a student eliminates some responses and ends up guessing between the two remaining responses, the student will sometimes get the right answer and sometimes the wrong answer. On average, the student will, by chance, answer half of these test items correctly, and so get some credit for partial knowledge.
1829	"[Doug] Lemov has his facts wrong: test score predict success in college only modestly, and they have very little predictive power after one takes high school grades into account. Decades of studies have shown this to be true of college admissions tests…"	College admission tests have substantial predictive power after controlling for other factors. That's why colleges use them; they don't have to. Admission test scores have about the same predictive power a high school GPA overall, but more at the top of the score distribution, where precision is very importan for many colleges. Moreover, the extra predictive power that is unique to college admission tests comes at a relatively very low price.

1829	"… and a few more recent studies have shown that scores on states' high-stakes tests don't predict better."	State's high-stakes tests are not designed to be predictive of college performance. They are focused retrospectively on mastery of the high school curriculum, which is important on its own. For example, civics knowledge may not predict college performance well but, as a society, we would like our students to master it in order to participate as good citizens.
1913	"But, in fact, despite all the care that goes into creating them [performance standards, cut scores, passing scores], these standards are anything but solid. They are arbitrary, and the "percent proficient" is a very slippery number."	Performance standards are subjective, not arbitrary The same is true for teachers' grading. The teacher, not some natural law, decides what level of performance deserves an A and what level of performance deserves an F.
2196	"They and their principal knew that the cost of forgoing test prep was often lower scores."	The preponderance of evidence indicates that inappropriate test prep—the type that Koretz and the Princeton Review claim increases scores—does not, on average, increase scores. A decade ago, the Better Business Bureau admonished the Princeton Review for false advertising.
2573	"…putting a stop to this disdain for evidence--this arrogant assumption that we know so much that we don't have to bother evaluating our ideas before imposing them on teachers and students--is one of the most important changes we have to make."	There may be no on earth holding more disdain for the evidence for the effects of testing than Daniel Koretz. He acknowledges only that tiny bit of the research literature that supports his biases. If he is not willing to change his close-minded ways, why should he expect anyone else to?
2717	"One reason we know less than we should … is that most of the abundant test score data available to us are too vulnerable to score inflation to be trusted. There is a second reason for the dearth of information, the blame for which lies squarely on the shoulders of many of the reformers."	Bunk. There exists an abundance of information. Th "dearth of information" exists only as a wishful dream inside Daniel Koretz's brain, and as a motivating force for his work. There may be no one in our country who has done more to dismiss and suppress the majority of the relevant research literature on education testing."
2757	"High-quality evaluations of the test-based reforms aren't common, …"	Actually, high-quality evaluation of testing interventions have been numerous and common over the past century. Most of them do not produce the results that Koretz prefers, however, so he declares them nonexistent.

2851	"It's no exaggeration to say that the costs of test-based accountability have been huge.'	Yes, it is an exaggeration. The cost of testing is close to inconsequential. The U.S. has instituted some sub optimal or ineffective accountability schemes (e.g., No Child Left Behind, Value-added Measurement teacher accountability). But, without any accountability structure, we arrive back at the Lake Wobegon that John J. Cannell, M.D. discovered in the 1980s.
2920	"The NRC panel was intentionally designed to include both supporters and opponents of test-based accountability. (I was a member of the NRC study panel.) The NRC panel concluded that test-based accountability systems had a modest positive impact on some aspects of student learning."	The US National Research Council's Board on Testin was captured by education establishment diehards in the late 1980s, and has tolerated no dissent since NRC panels on testing are carefully constructed shams, staffed with knowledgeable, manipulative insider experts who oppose testing and naive, relatively uninformed and easily manipulated testin supporter-novices.
3005	"As you know, experts in measurement have been warning for well over half a century that standardized tests can only measure a portion of the domains they target, such as math."	Do any tests exist that "target" all of math? Despite what Koretz wishes you to believe, many measurement experts, and probably most, disagree with him. Again, standards-based tests cover entire domains, and they do it relatively easily. Rarely are those tests asked to cover ALL of mathematics. Mor typically, a standards-based test might be charged with covering all the standards in a single grade-leve of math in a single jurisdiction.
3184	"But the failure to evaluate the reforms also reflects a particular arrogance."	What hubris. Koretz steadfastly avoids any debate, any evaluation or criticism of his convoluted claims. To hear him tell it, no researcher in the world disagrees with him, or can produce any evidence that counters his assertions.

3229	"I've several times excoriated some of the reformers for assuming that whatever they dreamed up would work well without turning to actual evidence."	This is shameful. More than anyone else, Daniel Koretz has worked tirelessly to convince people tha the research literature on the effects of testing doe not exist. Indeed, in the early 2000s, when our national leaders most needed to learn the lessons learned from a century's worth of research on testing effects, many of them turned to Koretz. And he told them that the cornucopia of information dic not exist. Above all else, Daniel Koretz works to suppress information--to convince today's policymakers not to look at the robust, extant research literature, most of which disagrees with him.
3248	"Finland, the Netherlands, and Singapore … none of the three has an educational accountability system remotely like ours."	He's right. All three of these other countries take student accountability much more seriously than w do. Students face more, and more consequential, high-stakes decision points in their student careers.
3256	"Finland has no high-stakes testing at all other than matriculation exams at the end of high school."	Finland has plenty of high-stakes testing, aside from what Koretz will admit to. The other high-stakes testing is designed, administered, and scored locally but it is still high stakes.
3323	"All students in Singapore are given standardized tests, but they encounter truly high-stakes tests far less often than American students, at only a few points in their careers: at the end of primary school and after four years of secondary school."	Contrast this statement with his next one. In the next statement, Koretz proclaims, truthfully, that American students encounter very few tests with consequences--only in some states with high-schoo exit exams. In this statement he wants us to believe that American students are inundated with high-stakes tests.
3334	"Although some high-stakes tests--in particular high-school exit tests--have direct consequences for students, most don't."	Contrast this statement with his previous one. In th previous statement, Koretz suggests that American students are overcome with high-stakes tests--that make them throw up and bang their heads on their desks. Here he admits, truthfully, that many American students face no high-stakes tests at all, and those who do face then at the end of high school.
3705	"Test scores can be improved very rapidly--even in the space of only two or three years--if one turns a blind eye to fraudulent gains."	Or, more accurately, if one turns a blind eye to lax security, as Koretz would have us do.

3772	"The first solid study documenting score inflation was presented twenty-five years before I started writing this book."	There was nothing "solid" about Koretz's circa 1990 study. It was, as far as we know, conducted in a still unknown school district, with unknown tests, and with no controls for any aspect of test administration. Moreover, despite all the bias in his favor in the education journal world, he has never managed to get the study published in a peer-reviewed journal. It remains just a generic conference presentation.
3772	"The first study showing illusory improvement in achievement gaps--the largely bogus "Texas miracle"--was public shed only ten years after that."	Koretz refers to a study conducted by his CRESST colleagues--the "October Surprise" of the 2000 presidential election campaign. The Texas gains wer not illusory; between 1990 and 1998 Texas' NAEP scores improved more than any other state's, bar one. The study by Koretz's colleagues, however, wa misrepresented by its own authors. They performec hypothesis tests separately on each NAEP test (in each subject, at each grade level), then made statements about Texas education as a whole. To make valid conclusions about the state as a whole they should have pooled the results from all grades and subjects. Had they done so, their results would have been highly statistically significant.

[1] http://www.tandfonline.com/doi/full/10.1080/15305058.2011.602920

[2] https://www.aei.org/events/has-k-12-education-fallen-for-a-testing-charade/