Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 1: Beyond All Reason

"Test-based accountability has become an end in itself in American education, unmoored from clear thinking about what should be measured, how it should be measured, or how testing can fit into a rational plan for evaluating and improving our schools."

Standardized High-Stakes Testing ❯ Produces ❯ Winners and Losers

"The prescription that has been imposed on educators and children in response is seductively simple: Measure student performance using standardized tests and use those measurements to create incentives for higher performance. If we reward people for producing what we want, the logic goes, they will produce more of it. Schools will get better, and students will learn more."

- Must change attitude about the purpose of high-stakes testing
- Pressure to raise achievement as measured by testing pervades schools
- Many harmful effects of test-based accountability
- Evidence of the lack of clear thinking about what should be measured and how to measure it
- When purpose of the tests is to show growth by test scores, instruction emphasizes the limited components of curriculum that can be measured and devalues the importance of that which is not easy to measure.
- Teachers focus heavily on the tested segment of a domain - little time for analytical thinking, problem-solving, and other concepts that are time consuming and difficult to assess
- Must confront "honestly the failures that stare us in the face"

"It's time for us to switch prescriptions, to put in place accountability systems that encourage teachers to act in ways that we do want and that produce students who are more capable–not just higher scoring on a few tests but ***more knowledgeable, more able to learn on their own, more able to think critically, and therefore more successful, not only in their later work but also as citizens.*** *To do this, we have to start by confronting honestly the failures that stare us in the face."*

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 2: What *Is* a Test?

Much of what has gone wrong with testing stems from misunderstandings of testing, best illustrated by George W. Bush's claim during the questions surrounding NCLB:
"A reading comprehension test is a reading comprehension test. And a math test in the fourth grade—there's not many ways you can foul up a test. It's pretty easy to norm."

"Large-scale tests are typically used to estimate mastery of some large area of study, called a "domain" in the testing world. These may reflect a full year of work (algebra) or more (skills in reading and language arts developed over a period of years). 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."

Testing Experts ➔ Should NEVER use test scores precisely the way they are used now.

"Rather than sampling a small number of people to represent the population as pollsters do, the authors of tests sample a small amount of content to represent the larger domain. Most of the domain remains untested, just as most voters are not reached by pollsters."

"Testing simply can't carry the freight that has been piled onto it. The failure to understand this, or a willful decision to ignore it, can explain much of what has gone wrong."

Consequences of sampling:

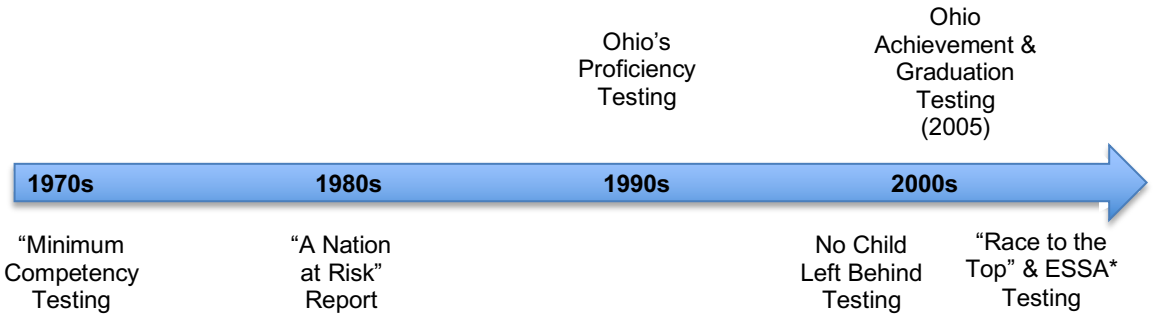
- Imprecision or "error"
  Tested samples of content and skills are not fully representative, either of the goals of schooling broadly or of student achievement more narrowly
- High-stakes testing creates strong incentives to focus on the tested sample rather than the domain it is intended to represent
- Perverse incentives are created for educators

"High-stakes testing creates strong incentives to focus on the tested sample rather than the domain it is intended to represent."

The result of the focus on a narrow part of the whole domain is bad test prep that robs students of useful instruction.

Info Yielded by Tests = Useful
<u>BUT Never Enough to Evaluate Programs – Schools – Educators</u>
Bottom Line

Chapter 3: The Evolution of Test-Based "Reform"



"…none of these different variations on the theme escapes the fundamental conclusion that this approach has been for the most part a failure."

Reasons for failure:

- System *"rewards far too narrow a slice of educational practice and outcomes."*
- System *"is very high pressure."*
- Reforms *"left almost no room for human judgment."*
- Absence of other incentives to balance pressure to raise scores.
- Every person in educational system is given same incentives."

"In the testing world, the increasingly high-stakes use of tests became known as 'measurement driven instruction.' Most jargon is worth forgetting, but this term is important because it signaled a fundamental change in the purpose of testing. Achievement testing had always been intended as a tool to improve instruction. The reforms didn't change that."

"However, in the traditional approach the main purpose of scores was to give teachers information that would help them teach more effectively. Improved scores would follow greater mastery of the curriculum, just as better polling results would follow an effective campaign. 'Measurement driven instruction' reversed this: tests would now lead. Improving performance on the specific task was to be the explicit goal, and higher quality instruction would be the consequence. This was the tail wagging the dog."

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 4: Campbell's Law

Donald Campbell, a pioneer of program evaluation in the social sciences, discovered that once an important measure or benchmark of performance is set for a system—e.g., manufacturing, health care, safety testing, education—the attention that is paid to this measure gives a false sense of improvement and usually damages other parts of the system not under such close scrutiny.

Campbell's Law – central to Koretz's contention that education reform efforts (No Child Left Behind, Every Student Succeeds, Common Core Tests) have resulted in the **illusion of improvement rather than genuine improvement**

"The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."
– Donald T. Campbell "Assessing the Impact of Planned Social Change. 1975"

Koretz's examples of non-education measurement of performance expectations gone awry:
- VW emission scandal, 2015
- Hospital ERs/Doctors' ratings
- Soviet manufacturing system

"Achievement tests may well be valuable indicators of... achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they lose their value as indicators of educational status and distort the educational process in undesirable ways."

Campbell's Law applies in any high-pressure accountability system that is based only on a few hard numbers.

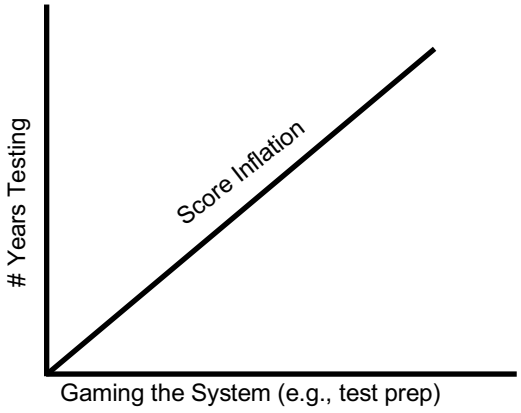
"Distortions" and "Corruption" produced by Campbell's Law
- test score inflation,
- gaming the system,
- outright cheating,
- unrealistic achievement targets,
- unfair teacher evaluations,
- disproportionate consequences for schools in poorer communities
- neglect of "important" educational experiences for ALL students

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 5: Score Inflation

> *"The entire logic of our reforms depends on rewarding the schools that do better and punishing those that don't. However, because in most contexts we can't separate score inflation from legitimate improvements, we are sometimes rewarding people who game the system more effectively, and we are punishing educators who do good work but appear to be doing relatively less well because they aren't taking as many shortcuts."*

- Evidence, accumulated over 25 years: **inflation common and often very large**
- Numerous studies have found that score gains were exaggerated three-to six fold, and there have been instances in which very large gains in scores have been accompanied by no real gains in learning
- Unwillingness of many people in public education to allow honest evaluations that risk findings they don't want – political risk aversion
- Koretz unaware of single credible study of high-stakes testing in U.S. that has failed to find inflation in at least some of the systems examined – pervasive
- Predictable sampling in test development leads to score inflation, e.g., content, format, items, scoring
- "Strategy of applied anxiety" – teachers feel strong pressure to raise scores used for evaluation
- Holding people accountable for unreachable targets leads many educators to take desperate measures
- Cannot separate score inflation from legitimate improvements, so we sometimes reward people who game the system effectively and punish educators not taking shortcuts
- Number of studies found that score inflation is **more severe among disadvantaged students**; schools with a higher proportion of poor students showed greater average inflation
- When targets require faster gains than teachers can produce by legitimate means, they have strong incentive to search for methods might raise scores quickly
- Primary focus is on accountability for educators, not students – hence gaming the system
- NAEP Scores not susceptible to inflation because teachers not held accountable for scores, hence no incentive to engage in NAEP-focused test prep
- Score inflation leads to emulating programs that look good only because of bogus score gains and overlooking programs that really are good because teachers using them engage in less gaming (aka false positives)



*"It's not surprising that **disadvantaged students suffer more from score inflation**... low performing schools often face severe barriers to improvement – for example, fewer resources, less experienced teaching staff, high rates of teacher turnover, higher rates of student transience, fewer high performing students to serve as models, fewer parents who are able to provide supplementary supports, and less pressure for academic achievement from parents, among other*

things. Faced with these obstacles, teachers will have a stronger incentive to look for shortcuts for raising scores."

Chapter 6: Cheating

Types of Cheating:
- Changing students' answers after-the-fact.
- Providing either teachers or students with test items in advance.
- Providing students with inappropriate assistance or giving them the answers.
- Excluding students who were likely to score poorly, i.e., "scrubbing."

"Cheating –by teachers and administrators, not by students–is one of the simplest ways to inflate scores, and if you aren't caught, it's the most dependable."

Cheating scandals usually show implausible score gains and a large number of wrong-to-right erasures.

It is difficult to verify cheating because the resistance of the system to a thorough and honest investigation can be difficult.

Individual **teachers** simply provide the illicit information to students directly, such as reviewing the test items in advance and preparing a study guide for students that includes exact items from the test items. Cheating is clearly widespread.

36%: said that test questions are occasionally or frequently rephrased during testing time
21%: said the same of answering questions about test content or recommending revisions of students' answers
17%: said that hints were given occasionally or frequently
9%: said that answers were changed at least occasionally

Why are we so unlikely to ever find out how common cheating has become – gullibility and willingness to accept seeming good news at face value.

-Score increases seem to confirm that the reforms are delivering exactly what was promised. This is a balloon that people are reluctant to pop.
-Second problem is simply the scale of the possible problem and resistance to thorough investigation – warning signs ignored for years
"In what may be the most cited academic study of cheating, Brian Jacob and Steven Levitt, using data from Chicago, estimated that 'serious cases of teacher or administrator cheating on standardized tests occur in a minimum of 4 –5% of elementary school classrooms annually.' However, they noted that their method of estimating cheating, which relies on unexpected fluctuations in scores and unusual answer patterns, is likely to underestimate the true prevalence

Who's responsible?
- Teachers and/or administrators who cheated responded to extreme pressure to raise scores?
- Is it just people who actually carry out the fraud or encourage it?
- Or, are those who create the pressures to cheat also culpable, even if not criminally?

because it does not detect some methods of cheating."

Chapter 7: Test Prep

"People don't agree on the dividing line between test prep and regular instruction…. But observe schools or talk to teachers or parents, and it's clear that test prep now absorbs a good bit of available time."

When Does Test Prep Become Cheating?

- Is some of this test prep dishonest or unethical?
- Is it really cheating?
- Where should we draw line between undesirable test prep and cheating?
- When teachers omit material they know is important for student success?
- Should test prep techniques that produced fraudulent gains be considered cheating?

Types of Bad Test Prep:

- Reallocation
  - between subjects – cut back on content not tested & shift resources to tested subjects
  - within a subject – focus time and other resources within tested subjects, focusing on content that is emphasized by the test.
- Coaching
  - focus on unimportant details of the particular test, i.e., format of the test items, how student responses are scored
  - "Pythagorean Triples"
  - Memorizing arbitrary symbols

Dissonance Reduction – cognitive dissonance is discomfort people feel when they hold two contradictory beliefs or values; people will sometimes revise what they think to reduce the contradiction.

"Not only is bad test prep pervasive. It has begun to undermine the very notion of good instruction."

Some young teachers have a hard time envisioning what instruction would look like without it – it's the new normal

For many teachers, raising scores had become the end goal, the mark of a "good" teacher. To an alarming degree, they had been taught that test prep and good instruction are the same thing.

Inappropriate test preparation – **more severe problem in schools serving high concentrations of disadvantage students**

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 8: Making Up Unrealistic Targets

- Reformers wanted testing to reduce inequities in the American educational system – failed miserably
- Establishing performance standards is more difficult than many supposed – results have been that standards almost always arbitrary and sometimes capricious
- Translation of standards into targets for teachers has created hugely negative consequences for students
- NCLB set targets that could never be reached

Reformers' Hubris

- Faced with unrealistic and unreasonable targets some educators cut corners or simply cheated – alternative was to stand by and watch disproportionate numbers of disadvantaged students fail
- Setting unrealistic targets cannot improve learning
- Standardized tests pretend that all kids are the same

"For decades, one of the primary–and most praiseworthy–goals of the test based reforms has been to reduce the glaring inequities in the American education system... part of the blame for this failure lies with the crude and unrealistic methods used to confront inequity. In a nutshell, the core of the approach has been simply to set an arbitrary performance target (the "proficient" standard) and declare that all schools must make all students reach it in an equally arbitrary amount of time. No one checked to make sure the targets were practical."

> *"If one doesn't look too closely, reporting what percentage of students are "proficient" seems clear enough. Someone somehow determined what level of achievement we should expect at any given grade–that's what we will call "proficient"–and we are just counting how many kids have reached that point... For the most part, the press reports differences among schools and progress over time only in terms of this single statistic... This trust in performance standards, however, is misplaced."*

"The primary motivation for setting a 'proficient' standard is to prod schools to improve, but information about how quickly teachers actually can improve student learning doesn't play much, if any, of a role in setting performance standards. When panels set standards, they are not given information about practical rates of improvement, and for the most part

they are not asked to consider them. They are just asked to try to figure out what level of performance constitutes proficiency."

Chapter 9: Evaluating Teachers

Tests fail to measure so many of the important goals of education – worse, emphasizing them at the expense of other important goals has created serious negative effects.
There should be a system that encourages teachers to improve and to weed out those who should not teach

Using student test scores in teacher evaluation – *"simplistic and unworkable"*

It *"defies logic to use tests to evaluate teachers."*

The Incompleteness of Tests – standardized tests cannot and do not indicate whether students "master" discipline areas; they ignore important things like engagement and curiosity.

Taking Test Scores Out of Context – reformers wanted measures could be used to evaluate teachers and schools without ever looking at the school from which they were obtained

It is a mistake to attribute scores to *"the actions of educators, despite the many other factors that influence student achievement."*

Value-Added Modeling (VAM) models – attempt to use earlier test scores and in some cases to predict student test scores.

"The estimate of a teacher's value added is obtained by adding deviation from predictions for all her students. If a teacher's students do better than predicted, it is taken to show that she is effective, but if they do worse she is ineffective."

Statistical Association of America: *"VAMs typically measure correlation, not causation: Effects—positive or negative—attributed to a teacher may actually be caused by other factors that are not captured in the model."*

Rating Students with the Wrong Test – if tests and the curriculum are not aligned, scores in no way indicate teacher effects

Teachers' Ratings are Inconsistent across Tests – estimates of teacher effectiveness can be highly sensitive to how the test samples from the domain

Teachers' Test Scores Are Unstable over Time: *"Despite the ongoing vehement arguments between opponents and supporters of using VAM, they generally agree that the estimates are very unstable over time."*

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

"However valuable tests may be for helping to evaluate schools and teachers, they can never be sufficient because they failed to measure so many of the important goals of education."

Chapter 10: Will the Common Core Fix This?

> *"The evolution of test-based accountability has shown a dreary sameness over the years and the Common Core, unfortunately, fits the same pattern to a T."*

_"No"_ the Common Core will <u>not</u> provide a better system for school accountability:

'Reformers' Hubris"

- Major flaw in the Common Core initiative was the <u>lack of field-testing</u>.
- Common Core test developers did not address a major cause of score inflation associated with using test item templates.
- Common Core uses a **"one-size fits all"** approach that sets arbitrary achievement targets based on college entrance standards and ignores inevitable variations of student populations and school characteristics.
- Common Core initiative is still too dependent on standardized test results.
- Pressure from Common Core is, in fact, increased because content has been chosen to be more "rigorous."

A new flavor of the same old thing

Test-based accountability:
- Stress
- Degraded instruction
- Bad test prep
- Score inflation
- Fraud/cheating
- Narrowed curriculum

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 11: Did Kids Learn More?

"What did we get in return for all of the stress, degraded instruction, bad test prep, score inflation, and outright fraud that test based accountability has engendered? Did students actually learn more? Yes and no."

"The honest answer is that it's very difficult to pin down with precision any effects the reforms had on actual student learning."

Test score data available to us are too vulnerable to score inflation to be trusted.
Reformers imposed system on students and teachers without taking time to evaluate testing programs

Making Sense of the Evidence

Two important questions to ask:

- What happened to student learning during the time of test based accountability?
- Why did such changes happen, and specifically whether they can be attributed to test based accountability?

Did Learning Improve?

Reading

- Answer is simple – trend data show that student learning has not improved much, despite pressure to raise test scores

Math

- Scores show impressive math gains of 4th graders, but they don't persist – wither as students progress thru grades

Trends in Achievement Gaps

> *"While the gap between black and white has been shrinking radically, the gap between rich and poor students (measured as the gap between students from families at the 10^{th} and 90^{th} percentile in income) has been widening consistently."*

How Much Did the Reforms Contribute to Trends in Achievement?

> *"These data make it clear that we haven't ended up even close to where the reformers wanted us to be, but they don't answer the harder question: just how much did test based accountability affect these trends?"*

Putting the Pieces Together

> *"It's no exaggeration to say that the costs of test based accountability have been huge. Instruction has been corrupted on a broad scale. Large amounts of instructional time are now siphoned off into test prep activities that at best waste time and at worst defraud students and their parents. Cheating has become widespread. The public has been deceived into thinking that achievement has dramatically improved and that achievement gaps have narrowed. Many students are subjected to severe stress, not only during testing but also for long periods leading up to it. Educators have been evaluated in misleading and in some cases utterly absurd ways. Careers have been*

disrupted and in some cases ended. Educators, including prominent administrators, have been indicted and even imprisoned."

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 12: Nine Principles for Doing Better

Advocates of test-based accountability started with good intentions, i.e., to address inequities and shortcomings in the American educational system.

> **The fundamental unwarranted assumption of test-based accountability was that performance as measured by standardized tests was sufficient to measure the quality of schools and/or teachers.**

Nine principles for doing better and correcting what has gone wrong:

Insanity is defined as doing the same thing over and over and expecting different results.

Pay Attention to Other Important Stuff

- Unwarranted assumptions that student achievement in a modest number of subjects covers enough of what we want from schools & that standardized tests are sufficient measure of that portion
- Cannot infer that differences in test scores or VAM correspond to school quality

> **We must decide what we want most to see in schools and then design systems to encourage it, not punish educators.**

Monitor More than Student Achievement

- Measuring schools effectively requires more than measuring just student achievement.

Set Reasonable Targets

- Set targets that the majority of educators can reach by legitimate means.

Stop Just Kicking the Dog Harder

- Sanctions and rewards are not enough
- Some teachers don't have supports necessary for success

Don't Expect Schools to Do it All

- Equity is required to increase achievement (pre-schools, wrap-around services, nutrition, health care, safety, etc.)

Pay Attention to Context

- *"Reformers who pushed test-based accountability believed that schools can be evaluated without anyone ever actually looking at them."*
- Must look at why students don't perform up to expectations

Accept the Need for Human Judgment

- Many reformers simply do not trust educators to evaluate schools
- Restore trust by valuing professional judgment

Create Counterbalancing Incentives

- Test-based accountability = exact same incentives for everyone – raise scores & not worry about how
- Educators need incentives NOT to behave badly

Monitor, Evaluate, and Revise

- Test-based accountability "reforms" were not evaluated and revised
- We need to evaluate what we do to avoid Campbell's law

> ***"It would be hard to justify continuing with an approach that does so much damage while creating so little benefit."***

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 13: Doing Better

Many "reformers" assumed that whatever they implemented would work well without turning to actual evidence.

Nothing justifies standing pat and continuing what we've been doing.

We must monitor whatever changes we make to the accountability system and be ready to make midcourse corrections.

Nations that rank highly in PISA and TIMMS – Finland, the Netherlands, and Singapore – have very different systems from ours – they use local measures of student achievement – trust educators.

Three themes:
- breadth (to combat Campbell's law)
- tradeoffs (no panaceas, tough choices)
- balance (create counterbalancing incentives)

Must Measure What Matters Most

The Big Three:
- student achievement
- educators' practice
- classroom climate

Problem has not been testing itself, but the misuse of testing

In addition to Big 3 we need to assess "soft" non-cognitive skills that cannot be measured on standardized tests

Give a **substantial role to the judgment of professionals**

Build sensible accountability system that measures a broad range of important things

Create a curriculum with reasonable targets that measure student growth, not level of their performance

Use tests sensibly – measure a broad range of important things – one test cannot do everything

Reduce or eliminate the use of "interim" or "benchmark" tests (e.g., MAP Tests)

Provide better supports for teachers, including better pre-service training, in-service training, in-school supports for students, & out of school supports such as preschool

Monitor the system and make midcourse corrections when necessary – field test approaches

Daniel Koretz, The Testing Charade, Pretending to Make Schools Better (2017)

Chapter 14: Wrapping Up

Everything that has happened with test-based accountability could have been predicted based on what was known 30 years ago

What wasn't anticipated was the intensity of the **subsequent damage**:

- "jaw-dropping" score inflation
- corruption of "the notion of good teaching,"
- indefensible methods of teacher evaluation

More flexible federal law (ESSA) and changing parental attitudes about testing (OPT OUT) is sign of a possible shift in the test-based accountability movement

Moving away from test-based accountability requires **"humility"** – better accountability systems:

- have to function in complex environments
- will be more expensive
- involve tradeoffs
- include some ideas that fail

There must be a long-term commitment to "monitor, reject, and revise," when ideas don't work or the effects of Campbell's Law become evident.

"Even though ESSA won't in itself do enough to reduce the distortions created by test-based accountability, this dissatisfaction with the past offers some hope that ESSA represents the beginning of a shift to a more sensible and productive approach . . . ***Many parents have become fed up with having their children in schools that are so dominated by testing.****"*

"In an important sense educators didn't fail. Teachers and principals didn't manage to make the improvements in education that the policy makers claimed, but they did precisely what was demanded of them: they raised scores."

> *"It's remarkable that even Arne Duncan, who arguably did as much as any one person during the past decade to increase the pressure on educators to raise test scores, conceded that 'testing issues today are sucking the oxygen out of the room in a lot of schools.'"*

"Will it be difficult to implement these suggestions? Yes, very, and expensive as well. Is there room to argue about how best to put them into practice? A great deal, and we will undoubtedly make some mistakes regardless of who wins those debates. And progress won't be fast; it will take quite some time simply to repair the damage that test based accountability has produced, let alone to make the sizable improvements we want. But years of experience have shown that the alternative–Dodging these difficulties and tinkering with what we have–is unacceptable."