Heights Coalition for Public Education 2019 Community Book Study
Summary: Daniel Koretz "The Testing Charade: Pretending to Make Schools Better"

Chapter #1: BEYOND ALL REASON

"The prescription that has been imposed on educators and children in response is seductively simple: Measure student performance using standardized tests and 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."(6)

"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."(9)

Chapter #2: WHAT IS A TEST?

"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." (13)

"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."(13)

"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."(15)

- 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

Chapter #3: THE EVOLUTION OF TEST-BASED "REFORM"

1970s: "Minimum competency" testing
1980s: "A Nation A Risk" report
1990s: "No Child Left Behind" NCLB, "measurement based" testing
2000s: "Every Student Succeeds Act"

- Explicitly acknowledges that scores alone are insufficient.
- States must include graduation rates for high schools.
- States are required to include one additional measure of school quality or student success: student engagement, educator engagement, student access to and completion of advanced course work, post secondary readiness, school climate and safety, attendance.

"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."(26)

"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."(26)

Potential causes of reform failure:

- The system it imposed on schools rewards far too narrow a slice of educational practice and outcomes.
- The system is very high-pressure.
- It left almost no room for human judgment. Teachers are not trusted to evaluate students or each other, principals are not trusted to evaluate teachers, and the judgment of professionals from outside the school has only a limited role.
- The system lacks any other incentives to balance the pressure to raise scores. Absolutely no one is given any incentive to monitor or control **how** these gains are achieved.

Chapter #4: CAMPBELL'S LAW

"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."(38)

"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." (39)

Important lessons learned from Campbell's Law:

- It will show up in any high-pressure accountability system that is based only on a few hard numbers.
- We need to look at the net effect of reforms, balancing the negative effects against the positive.
- In educational testing the corruption of measures about which Campbell warned takes the form of ***score inflation***– increases in scores larger than improvements in learning justify:
    - Cheating
    - Reallocation
    - Excluding people with bad numbers
    - Lowered standards

Heights Coalition for Public Education 2019 Community Book Study
Summary: Daniel Koretz "The Testing Charade: Pretending to Make Schools Better"

Chapter #5: SCORE INFLATION

National Assessment of Educational Progress (NAEP) – a set of standardized tests that are widely considered a gold standard for evaluating educational trends.

- 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. And NAEP scores are there for the taking. In math and reading, NAEP is administered every two years, and the scores are available to anyone on the web. (57)

- Joel Klein: Chancellor, New York City Schools (2002-2010), David Steiner: New York State Commissioner of Education, Meryl Tisch: New York State Chancellor of Board of Regents.
    - The state tests showed improvement of 8th grade math students, but showed no improvements on the NAEP. New York test scores were inflated.

"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. On top of that, we are holding out as examples to be emulated programs that look good only because of bogus score gains and overlooking programs that really *are* good because the teachers using them are doing less to game the system. In other words, the system can propagate a bad practice."(64)

"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."(68)

Chapter #6: CHEATING

"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." (73)

How Do People Cheat?

- 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, a technique that is often called "scrubbing".

Cheating Scandals

"The most important lesson… is that holding people accountable for reaching targets that they can't reach by legitimate means has led many educators to take desperate measures."(79)

- Atlanta
- Philadelphia
- Washington DC

Cheating by Individual Teachers

"The press has documented a number of cases in which teachers or principals opened test booklets in advance. In some cases principals gave this information to teachers so that they could prepare students for the specifics of the test."(87)

"One indication that these teachers knew full well that they were cheating was the creative ways some of them tried to mask what they were doing– for example using facial expressions, hand gestures, and pre-established code words to communicate that student answers were right or wrong, rather than simply telling students out right."(87)

How Common Has Cheating Become?

"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 because it does not detect some methods of cheating." (88)

Apportioning Blame

- Who is responsible for cheating? And who is punished?
- This prompts the larger and more uncomfortable question: Just who is responsible?
    - Is it just the people who actually carry out the fraud or require it?
    - Or are those who create the pressures to cheat also culpable, even if not criminally?

Chapter #7: TEST PREP

"People don't agree on the dividing line between test prep and regular instruction. In fact, one of the most pernicious effects of reform has been to blur– and in some cases entirely obscure– the distinction between test prep and teaching. But observe schools or talk to teachers or parents, and it's clear that test prep now absorbs a good bit of available time."(93)

"Of course there is good test preparation. One of the legitimate purposes of tests is to help teachers learn what aspects of their teaching need to be changed or strengthened, which ultimately benefits their students."(94)

Heights Coalition for Public Education 2019 Community Book Study
Summary: Daniel Koretz "The Testing Charade: Pretending to Make Schools Better"

Three Types of Bad Test Prep

- Reallocation between subjects
  - Cut back on things that don't count and shift resources to the tested subjects.
- Reallocation with in a subject
  - Reallocating time and other resources within tested subjects, focusing on content that is emphasized by the test.
- Coaching
  - Focuses on unimportant details of the particular test, such as the format of the test items, other aspects of the presentation of material, and how student responses are scored.
  - "Pythagorean Triples"
  - Memorizing arbitrary symbols

When Does Test Prep Become Cheating?

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

"Psychologists have a term for this: *dissonance reduction*. *Cognitive dissonance* refers to the discomfort people feel when they hold two contradictory beliefs or values. To reduce the stress, people will sometimes revise what they think to reduce the contradiction."(112)

Corrupting the Idea of Good Teaching

"Not only is bad test prep pervasive. It has begun to undermine the very notion of good instruction. This has happened in part just because of the passage of time. High-stakes testing and undesirable test prep have been in place for so long that many young teachers have spent their entire careers immersed in them. As some young teachers have told me, they simply have a hard time envisioning what instruction would look like without it."(112)

"They [new teachers] were telling me that I was missing the boat by seeing test prep as something that competes for time with good instruction. In their experience, *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."(113)

And What About Equity?

"Inappropriate test preparation, like score inflation, is more severe in some places than in others… So one would expect that test preparation would be a more severe problem in schools serving high concentrations of disadvantage students, and it is. Once again, disadvantaged kids are getting the short end of the stick. Ironically, some aspects of the reforms that were intended to help disadvantaged students appear to have contributed to this demoralizing result."(117)

Chapter #8: MAKING UP UNREALISTIC TARGETS

"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."(120)

Making Up Targets

"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."(120)

"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."(125)

Pretending That All Kids Are The Same

"How much variation among kids should we expect? Just how much can we shrink this variation? Short of cheating or inflating scores in other ways, there are only two ways to bring very low-scoring students to a high standard: dramatically increase everyone's scores or dramatically increase just the performance of kids at the bottom."(130 – 131)

Chapter #9: EVALUATING TEACHERS

"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. Worse, emphasizing them at the expense of other important goals had created some very serious negative effects."(137)

"I told Duncan and his staff that this is a fine example of 'other important stuff'–making math interesting and even fun, keeping kids engaged, getting students to reason about math rather than simply practicing procedures, teaching kids to communicate about their work, and helping students to focus more on learning than on simply being right. This is precisely the sort of thing I hope to see when I walk into a classroom. And I want it for all kids, not just my own."(139)

Why Test Scores Provide Such A Problematic Measure of Educator Performance

- The incompleteness of tests
    - There are important aspects of the mastery of mathematics, for example, that we can't capture well– or at all– with current tests.
- Taking the test scores out of context

  - Reformers wanted measures that someone sitting in a state capital could interpret without ever looking at the school from which they were obtained.
  - Test scores taken out of context often don't tell you what you need to know.
- Trying to use tests to explain, not just describe
  - Mistakenly attributing scores – high or low – to the actions of educators, despite the many other factors that influence student achievement.
  - Of course the actions of educators do affect scores, but so do many other factors both inside and outside of school, such as their parents' education.
- Using "Value Added Modeling" (VAM) to evaluate teachers
  - The American Statistical Association statement on using value added models for educational assessment reads: "VAMs typically measure a 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." "That is, if a VAM is used to estimate your 'effectiveness' as a teacher, that estimate will sometimes blame or credit you for things that have nothing to do with your teaching." (151)
- Rating teachers with the wrong test.
- Teachers' ratings are inconsistent across tests.
- Teachers' test scores are unstable over time.
  - Reliability and bias

Chapter #10: WILL THE COMMON CORE FIX THIS?

"No" (161)

- Predictability
- One-size-fits-all
- Over reliance on test scores
- Excessive pressure
- A new flavor of the same old thing

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."(175)

"The honest answer is that it's very difficult to pin down with precision any affects the reforms had on actual student learning."(175)

- Most of the abundant test score data available to us are too vulnerable to score inflation to be trusted.
- The reformers declared that they had figured out what would work, and they imposed it on students and teachers on a mass scale without taking time to evaluate their programs first. (175)
- They also often implemented their programs in a way that made it harder for anyone else to evaluate them. We normally evaluate programs by comparing people who participate in them to similar people who don't, and that requires that we keep some people out of the program long enough to make this comparison. (176)

Heights Coalition for Public Education 2019 Community Book Study
Summary: Daniel Koretz "The Testing Charade: Pretending to Make Schools Better"

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?
    - Simple trends on tests like Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA), and National Assessment of Educational Progress (NAEP) reflect everything that can affect scores, not just the reforms.
    - Are scores going up more than they would be in the absence of the reforms?
    - Are scores going down less rapidly than they would be?

Did Learning Improve?

- Reading
  - In the case of reading, the answer is simple: trend data show that student learning hasn't improved much, despite decades of unrelenting pressure to raise test scores in reading. (178)
- Math
  - While scores show impressive math gains of fourth graders, they don't persist: they wither as students progress through school.

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 10th and 90th percentile in income) has been *widening* consistently."(187)

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?"(189)

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." (191)

Chapter #12: NINE PRINCIPLES FOR DOING BETTER

1. Pay attention to other important stuff.
2. Monitor more than student achievement.
3. Set reasonable targets.
4. Stop just kicking the dog harder.
5. Don't expect schools to do it all.
6. Pay attention to context.
7. Accept the need for human judgment.
8. Create counterbalancing incentives.
9. Monitor, evaluate, and revise.

Chapter #13: DOING BETTER

How Is It Done Elsewhere? Systems Very Unlike Ours?

- Finland
    - No high-stakes testing, one matriculation exam after high school
    - Virtually no private schools
    - Teaching is highly regarded and well-paid.
- The Netherlands
    - Use of the various standardized testing
    - Public and private schools that are evaluated by the government but created by the citizens- market driven
    - Teaching is highly regarded and well-paid.
- Singapore
    - Highly centralized educational system
    - Use of standardized test, but far fewer than an American students
    - Students have a familial and cultural obligation to do well on the tests.
    - Government is now realizing that the educational system needs to be more student centered rather than test driven.

Options For Doing Better

1. We must measure what matters: *The Big Three*.
    a. Student achievement
    b. Educators' practices
    c. Classroom climate
2. We need to measure *The Big Three* well.
    a. Disagreements will arise in deciding **how** to measure them.
3. We must build a sensible accountability system.
    a. The system has to emphasize what is important.
    b. Create counterbalancing incentives.
    c. Look well beyond what happens on any single day in the classroom.
    d. We need measures that are not too closely aligned with each other– That is, that are not too similar.
    e. Targets have to be reasonable: the goals facing educators have to be ones that they can reach buy legitimate means.
        i. Set goals based on students' growth, not the level of their performance.

   ii. Targets for growth can be made more stringent for low achieving students, as long as they remain practical and needed supports are provided.
4. Use tests sensibly.
   a. Make them less predictable in order to whittle away at inappropriate test prep.
   b. Stop pretending that one test can do everything.
   c. Curtail sharply the use of "interim" or "benchmark" assessments that are widely used to predict how students will score at the end of the year.
   d. Treat scores as the starting point rather than the endpoint of evaluation.
5. Provide support to teachers.
   a. Better initial training
   b. In school supports
      i. Supplementary classes
      ii. Longer school days
      iii. Smaller classes
   c. Out of school supports
      i. High-quality preschool
6. Monitor and make midcourse corrections.
   a. Test out new approaches before we take them to scale.
   b. Monitor and evaluate these approaches– routinely– once they are put in place.
   c. Make midcourse corrections when problems are uncovered.

<u>Chapter #14 WRAPPING UP</u>

"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."(244)

"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.'" (246)

"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."(248)