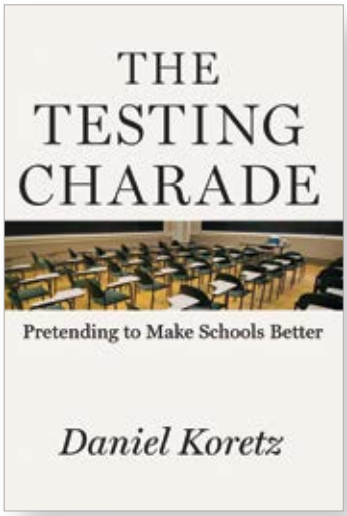
Needs Improvement

A gloomy perspective on high-stakes testing

The Testing Charade: Pretending to Make Schools Better
by Daniel Koretz

University of Chicago Press, 2017, $25; 288 pages.

As reviewed by Dan Goldhaber



The title of his latest book telegraphs where Harvard education professor Daniel Koretz stands on one of today's most contentious schooling issues: high-stakes testing. In short, this book is about "the failures of test-based accountability."

Koretz is an expert on testing and related policy, and his knowledge shines through in the book's early chapters, in which he discusses what tests are—and importantly, what they aren't. In particular, Koretz reminds us that because we cannot test for everything, tests only capture a slice of the academic and other skills we expect schools to help students master. Koretz also reminds us of the history of testing as a policy tool: test-based accountability long predates the 2001 passage of No Child Left Behind (NCLB), suggesting that the passage of NCLB's successor in 2015, the Every Student Succeeds Act, is unlikely to eliminate it.

The book gets to the heart of the matter in Chapter 4 on "Campbell's Law." This principle, penned by the social scientist Donald T. Campbell in 1976, suggests that "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." Koretz uses this precept to frame the discussion of test-based accountability efforts as well as more in-depth discussion in later chapters about some of the more pernicious corruptions of test-based accountability: artificial test-score inflation, undesirable types of test preparation, and outright cheating.

Koretz shows that using tests to hold schools and educators accountable for student achievement can lead to behaviors that don't support genuine learning. But here the portrayal feels slanted. For instance, Koretz states that "cheating has become a widespread scourge in our schools." Yet he concedes there is no way to know the prevalence of cheating, and notes that a well-cited study on cheating's prevalence in the Chicago public schools suggests that it occurs in 4 to 5 percent of elementary classrooms annually (see "To Catch a Cheat," *research,* Winter 2004). Indeed, it is somewhat ironic that Koretz points to the District of Columbia Public Schools (DCPS) as a prime example of the scourge of cheating. It stands to reason that cheating might proliferate there: DCPS's IMPACT accountability system (which uses test scores and other measures) is one of the most high stakes in the country. However, a series of studies by Thomas Dee and James Wyckoff show fairly conclusively that IMPACT has had a positive effect on teacher quality (see "A Lasting Impact," *research,* Fall 2017). Moreover, DCPS students show impressive gains over the last decade, not only on district tests but also on National Assessment of Educational Progress (NAEP) reading and math assessments. The NAEP progress is particularly relevant here, since throughout the book, Koretz treats NAEP as the gold-standard "audit test," that is, one that's not subject to the kinds of manipulation he describes. Thus, while Koretz has reason to be concerned about the perils of test-based accountability, evidence from DCPS suggests that it can work—when "it" is a nuanced system that uses more than tests alone to evaluate schools and teachers (more on this below).

Tucked into the middle of the book, in a chapter about teacher evaluation, is a passage that gets to the crux of the debate on accountability:

> The failures of test-based accountability shouldn't blind us to the serious and extraordinarily difficult problem that reformers were trying to confront. It was abundantly clear that in most districts there was no effective accountability for teachers after they were granted tenure, which in most locations requires only a few years of teaching. . . . Teachers who weren't competent . . . were allowed to continue teaching and often didn't even face any intervention.

Koretz is right: we need to move beyond subjective measures of educator performance, because such systems often do not provide honest assessments. For the most part, however, his critiques of test-based accountability do not shed light on how non–test-based systems might confront that central dilemma. Also, there is a logic to using tests to evaluate teachers and schools, because test scores do predict later-life outcomes such as college-going and earnings; and important recent evidence from Stanford researcher Raj Chetty and colleagues shows that having a "high value-added" teacher—one who improves student test scores—also

positively predicts these outcomes.

In the last part of the book, Koretz offers thoughts about what we can learn from other countries and how we can do accountability better. This part of the book falls short. While the different ways that other high-achieving countries monitor school performance are interesting, it is far from clear that such initiatives would work in the United States, given the vast contextual differences. Much-touted Finland, for instance, uses a school inspectorate model that is based on professional judgment. But teaching in Finland is also a highly compensated and selective profession. Thus, it is likely that cultural norms and professional standards in schools are quite different there, potentially implying greater receptivity to accountability. Koretz concedes that we can't say with certainty that adopting alternative accountability policies here would work better than test-based accountability, which may leave readers wondering whether they would.

In the discussion of how we might improve our accountability practices, Koretz writes that "advocates of test-based accountability may argue that I am oversimplifying their ideas." I think he is right! His ideas for improvement—for example, "Pay attention to other important stuff," "Don't expect schools to do it all," "Accept the need for human judgment," "Set reasonable targets"—are juxtaposed against a caricature of what most of those advocating test-based accountability likely believe. That is, they probably would agree with Koretz's suggestions.

While Koretz has reason to be concerned about the perils of test-based accountability, evidence from the D.C. Public Schools suggests that it can work, when it uses more than tests alone.

These straw-man arguments bring me back to the discussion of DCPS, for while IMPACT does use tests, it by no means relies exclusively on them. Human judgment is very much part of its evaluation system, which itself is part of a larger effort focused on sound hiring practices and significant feedback and supports for educators. This system's success illustrates how Koretz paints test-based accountability with too broad a brush.

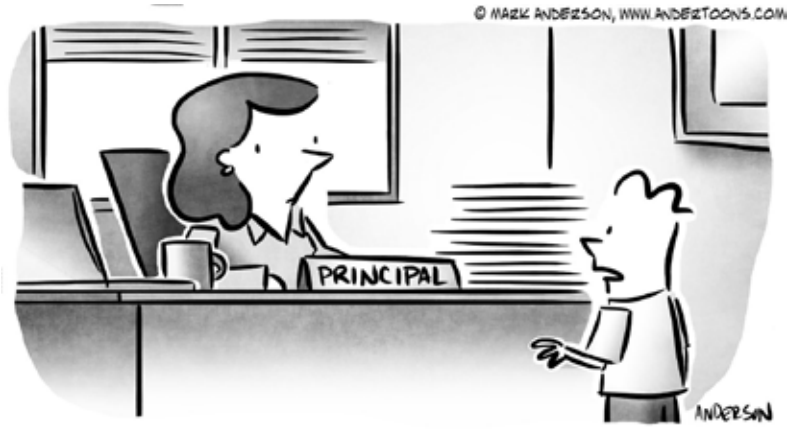
Koretz bluntly states that high-stakes testing has been "a failure" and that "almost thirty years before I started writing this book, I predicted that test-based accountability—then in its early stages, and still far milder than the system burdening schools today—wouldn't succeed. . . . I take no comfort in having been right."

But it is not clear from the evidence Koretz presents that he really was right. Indeed, in a chapter examining whether kids actually did learn more under test-based accountability, he asks: "What did we get in return for all the stress? . . . [There] are some bright spots, but the reforms didn't deliver the large gains in learning that would make us more competitive in international comparisons." In particular, since 2001 (that is, since NCLB was passed), there have been sizable gains in NAEP 4th- and 8th-grade math tests, small improvements in 4th- and 8th-grade reading tests, and very little change in 12th-grade scores.

These increases would seem to contradict the author's portrayal of test-based accountability as an unmitigated disaster. In fact, as Koretz acknowledges, the best cross-state study of NCLB (by Thomas Dee and Brian Jacob) suggests modest gains on the whole. It is reasonable, then, to wonder whether we need to abandon high-stakes testing altogether or whether better tests and smarter measurement of school and educator performance might help address the failings that Koretz describes.

What will readers take away from all this? Those who are already skeptical about using tests to judge schools and educators will find a lot to like, while those in favor of the practice will be challenged by the evidence presented that test-based accountability can lead schools to engage in unproductive practices. But readers who are new to the topic will not get a full picture of the failures and successes of test-based accountability—nor of the true extent to which Campbell's Law has (or hasn't) played out in this reform effort.

Dan Goldhaber is director of the Center for Education Data and Research at the University of Washington.



"You know, in the tech world being disruptive is seen as a positive."

CARTOON/©MARK ANDERSON