Book Reviews

Measuring Up: What Educational Testing Really Tells Us
by Daniel Koretz, Harvard University Press, 2008
Reviewed by Philip Staradamskis

In *Measuring Up*, Daniel Koretz continues his defense of the theory with which he is most famously identified: "Score inflation is a preoccupation of mine." He argues that high-stakes testing induces "teaching to the test," which in turn produces artificial test-score gains (i.e., test-score inflation). The result, according to Koretz:

> Scores on high-stakes tests—tests that have serious consequences for students or teachers—often become severely inflated. That is, gains in scores on these tests are often far larger than true gains in students' learning. Worse, this inflation is highly variable and unpredictable, so one cannot tell which school's scores are inflated and which are legitimate. (p. 131)

Thus, Koretz, a long-time associate of the federally funded Center for Research on Evaluation, Standards, & Student Testing (CRESST), provides the many educators predisposed to dislike high-stakes tests anyway a seemingly scientific (and seemingly not self-serving or ideological) argument for opposing them. Meanwhile, he provides policymakers a conundrum: if scores on high-stakes tests improve, likely they are meaningless—leaving them no external measure for school improvement. So they might just as well do nothing as bother doing anything.

Measuring Up supports this theory by ridiculing straw men—declaring a pittance of flawed supporting evidence sufficient (pp. 11, 59, 63, 132, and chapter 10) and a superabundance of contrary evidence nonexistent—and mostly by repeatedly insisting that he is right. (See, for example, chapter 1, pp. 131–133, and pp. 231–236.) He also shows

educational HORIZONS® *(ISSN 0013-175X) is published quarterly by Pi Lambda Theta, Inc., P.O. Box 6626, Bloomington, IN 47407-6626. Pi Lambda Theta membership includes a subscription to* **educational HORIZONS**®. *Nonmember subscriptions are available for $21 per year, U.S.; $32 per year, Canada and international. Periodicals postage paid at Bloomington, Indiana, and other mailing offices. Single copies: U.S. $6, Canada $6.50, and international $8, plus $1.75 postage. POSTMASTER: Address changes should be mailed to: Pi Lambda Theta, P.O. Box 6626, Bloomington, IN 47407-6626. All claims must be made within four months of publication. Back volumes available on microfilm from National Archive Publishing Co., 300 Zeeb Rd., P.O. Box 1346, Ann Arbor, MI 48106-1346; (800) 420-6272; outside N.A., (734) 302-6500; www.napubco.com. Indexed in* Current Index to Journals in Education *(ERIC). Selected back issues available online at www.pilambda.org.*

little patience for those who choose to disagree with him. They want "simple answers," speak "nonsense," assert "hogwash," employ "logical sleight[s] of hand," write "polemics," or are "social scientists who ought to know better."

Lake Wobegon

The concept of test-score inflation emerged in the late 1980s from the celebrated studies of the physician John J. Cannell (1987, 1989). Dr. Cannell caught every U.S. state bragging that its students' average scores on national norm-referenced tests were "above the national average," a mathematical impossibility. The phenomenon was dubbed the "Lake Wobegon Effect," in reference to the mythical radio comedy community in which "all the children are above average."

What had caused the Lake Wobegon Effect? Cannell identified several suspects, including educator dishonesty and conflict of interest; lax test security; and inadequate or outdated norms. But Cannell's seemingly straightforward conclusions did not make it unscathed into the educational literature. For instance, one prominent CRESST study provided a table with a cross-tabulation that summarized (allegedly all) the explanations provided for the spuriously high scores (Shepard 1990, 16). Conspicuously absent from the table, however, were Cannell's two primary suspects—educator dishonesty and lax test security.

Likewise, Koretz and several CRESST colleagues followed up with their own study in an unnamed school district, with unnamed tests and unidentified content frameworks. Contrasting a steadily increasing rise in scores on a new, "high stakes" test with the substantially lower scores recorded on an older, no-stakes test, Koretz and his colleagues attributed the inflation to the alleged high stakes.[1] Not examined was why two different tests, developed by two completely different groups of people under entirely separate conditions, using no common standard for content, would be expected to produce nearly identical scores.

This research framework presaged what was to come. The Lake Wobegon Effect continued to receive considerable attention, but Cannell's main points—that educator cheating was rampant and test security inadequate—were dismissed out of hand and persistently ignored thereafter. The educational consensus, supported by the work of CRESST and other researchers, fingered "teaching to the test" for the crime, manifestly under pressure from the high stakes of the tests.

Problematically, however, only one of Cannell's dozens of score-inflated tests had any stakes attached. All but that one were no-stakes diagnostic tests, administered without test-security protocols. The absence of security allowed education administrators to manipulate various aspects of the tests' administration, artificially inflate scores, and

then advertise the phony score trends as evidence of their own managerial prowess. Ironically, many of the same states simultaneously administered separate, genuinely high-stakes tests with tight security and no evidence of score inflation.

* * *

Much of *Measuring Up* recapitulates the author's earlier writings, but on page 243, we do learn what he and his colleagues actually found in that influential follow-up to Cannell's findings. Exactly why had scores risen so dramatically on the new, high-stakes third-grade test they examined?

> [A]lthough the testing system in this district was considered high-stakes by the standards of the late 1980s, by today's standards it was tame. There were no cash awards . . . [or] threats to dissolve schools or remove students in response to low scores. . . . The pressure arose only from less tangible things, such as publicity and jawboning.

In other words, this foundational study had involved no real high-stakes test at all. After all, in our open democracy, *all* tests are subject to "publicity and jawboning," whether they genuinely carry high stakes or no stakes. (Koretz, incidentally, is also incorrect in characterizing the test as "high stakes by the standards of the late 1980s": at the time more than twenty states administered high school graduation exams—for which failing students were denied diplomas.)

Do as I Say, Not as I Do

Many testing researchers (unsurprisingly, not associated with CRESST) caution against the simplistic assumptions that any test will generalize to any other simply because they have the same subject field name or that one test can be used to benchmark trends in the scores of another (Bhola, Impara, and Buckendahl 2003, 28; Impara 2001; Buckendahl et al. 2000; Impara et al. 2000; Plake et al. 2000; Archbald 1994; Cohen and Spillane 1993, 53; Freeman et al. 1983). Ironically, despite himself, Koretz cannot help agreeing with them. Much of the space in *Measuring Up* is devoted to cautioning the reader against doing exactly what he does—making apples-to-oranges comparisons with scores or score trends from different tests. For example:

> One sometimes disquieting consequence of the incompleteness of tests is that different tests often provide somewhat inconsistent results. (p. 10)

> Even a single test can provide varying results. Just as polls have a margin of error, so do achievement tests. Students who take more than one form of a test typically obtain different scores. (p. 11)

> Even well-designed tests will often provide substantially different views of trends because of differences in content and other aspects of the tests' design. . . . [W]e have to be careful not to place too much confidence in detailed findings, such as the precise size of changes over time or of differences between groups. (p. 92)
>
> [O]ne cannot give all the credit or blame to one factor . . . without investigating the impact of others. Many of the complex statistical models used in economics, sociology, epidemiology, and other sciences are efforts to take into account (or "control for") other factors that offer plausible alternative explanations of the observed data, and many apportion variation in the outcome—say, test scores—among various possible causes. . . . A hypothesis is only scientifically credible when the evidence gathered has ruled out plausible alternative explanations. (pp. 122–123)
>
> [A] simple correlation need not indicate that one of the factors causes the other. (p. 123)
>
> Any number of studies have shown the complexity of the noneducational factors that can affect achievement and test scores. (p. 129)

Recommendation Recoil

Koretz's vague suggestion that educators teach to "a broader domain" would dilute coverage of required content that typically has been developed through a painstaking public process of expert review and evaluation. In its place, educators would teach what, exactly? Content that Koretz and other anti-standards educators prefer? When the content domain of a test is the legally (or intellectually) mandated curriculum, teachers who "teach to the test" are not only teaching what they are told they should be teaching, they are also teaching what they are legally and ethically obligated to teach (Gardner 2008).

Another example of an imprudent recommendation: the *Princeton Review* sells test-preparation services, most prominently for the ACT and SAT college admission tests. Its publishers argue that students need not learn subject matter to do well on the tests, only learn some test-taking tricks. Pay a small fortune for one of their prep courses and you, too, can learn these tricks, they advertise. Curiously, independent studies have been unable to confirm the *Review*'s claims (see, for example, Camara 2008; Crocker 2005; Palmer 2002; Tuckman and Trimble 1997; Tuckman 1994), but Koretz supports them: "[T]his technique does often help to raise scores."

Scripting a Hoax

Around 1910, a laborer at the Piltdown quarries of southern England discovered the first of two skulls that appeared to represent the missing link between ape and human. In the decades following, mainstream science and some of the world's most celebrated scientists would accept "Piltdown man" as an authentic specimen of an early hominid. Along the way, other scientists, typically of the less-famous variety, proffered criticisms of the evidence, but they were routinely ignored. Only in the 1950s, after a new dating technique applied to the fossil remains found them to be modern, was the accumulated abundance of contrary evidence widely considered. The Piltdown fossils, it turned out, were cleverly disguised forgeries.

"Piltdown man is one of the most famous frauds in the history of science," writes Richard Harter in his review of the hoax literature (1996–1997). Why was it so successful? Harter offers these explanations:

- some of the world's most celebrated scientists supported it;
- it matched what prevailing theories at the time had led scientists to expect;
- various officials responsible for verification turned a blind eye;
- the forgers were knowledgeable and skilled in the art of deception;
- the evidence was accepted as sufficient despite an absence of critical details; and
- contrary evidence was repeatedly ignored or dismissed.

Measuring Up's high-stakes-cause-test-score-inflation mythmaking fits the hoax script perfectly.

References

Archbald, D. 1994. *On the Design and Purposes of State Curriculum Guides: A Comparison of Mathematics and Social Studies Guides from Four States* (RR-029). Consortium for Policy Research in Education.

Bhola, D. D., J. C. Impara, and C. W. Buckendahl. 2003. "Aligning Tests with States' Content Standards: Methods and Issues." *Educational Measurement: Issues and Practice* (Fall): 21–29.

Buckendahl, C. W., B. S. Plake, J. C. Impara, and P. M. Irwin. 2000. Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, La.

Camara, W. J. 2008. "College Admission Testing: Myths and Realities in an Age of Admissions Hype." In *Correcting Fallacies about Educational and Psychological Testing*, ed. R. P. Phelps, chapter 4. Washington, D.C.: American Psychological Association.

Cannell, J. J. 1987. *Nationally Normed Elementary Achievement Testing in*

America's Public Schools: How All Fifty States Are above the National Average. 2nd ed. Daniels, W. Va.: Friends for Education.

———. 1989. *How Public Educators Cheat on Standardized Achievement Tests*. Albuquerque, N.M.: Friends for Education.

Cohen, D. K., and J. P. Spillane. 1993. "Policy and Practice: The Relations between Governance and Instruction." In *Designing Coherent Education Policy: Improving the System,* ed. S. H. Fuhrman, 35–95. San Francisco: Jossey-Bass.

Crocker, L. 2005. "Teaching for the Test: How and Why Test Preparation Is Appropriate." In *Defending Standardized Testing,* ed. R. P. Phelps, 159–174. Mahwah, N.J.: Lawrence Erlbaum.

Freeman, D., et al. 1983. "Do Textbooks and Tests Define a National Curriculum in Elementary School Mathematics?" *Elementary School Journal* 83(5): 501–514.

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

Harter, R. 1996–1997. "Piltdown Man: The Bogus Bones Caper." *The TalkOrigins Archive*. Downloaded May 13, 2008, from <http://www.talkorigins.org/faqs/piltdown.html>.

Impara, J. C. 2001. Alignment: One element of an assessment's instructional utility. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, Wash., April.

Impara, J. C., B. S. Plake, and C. W. Buckendahl. 2000. The comparability of norm-referenced achievement tests as they align to Nebraska's language arts content standards. Paper presented at the Large Scale Assessment Conference, Snowbird, Utah, June.

Palmer, J. S. 2002. Performance Incentives, Teachers, and Students: Estimating the Effects of Rewards Policies on Classroom Practices and Student Performance. Ph.D. dissertation. Columbus, Ohio: Ohio State University.

Plake, B. S., C. W. Buckendahl, and J. C. Impara. 2000. A comparison of publishers' and teachers' perspectives on the alignment of norm-referenced tests to Nebraska's language arts content standards. Paper presented at the Large Scale Assessment Conference, Snowbird, Utah, June.

Shepard, L. A. 1990. "Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test?" *Educational Measurement: Issues and Practice* (Fall): 15–22.

Tuckman, B. W. 1994. Comparing incentive motivation to metacognitive strategy in its effect on achievement. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., April 4–8. Available from ERIC (ED368790).

Tuckman, B. W., and S. Trimble. 1997. Using tests as a performance incentive to motivate eighth-graders to study. Paper presented at the annual meeting of the American Psychological Association, Chicago, August. Available from ERIC (ED418785).

Note

1. The study traced the annual trend in average scores on a third-grade test "perceived to be high stakes" over several years, then administered a different third-grade test, with no stakes, that had been administered in the district several years earlier. The researchers, finding a steadily increasing rise in scores on the new test contrasted with a substantially lower score on the old, no-stakes test, attributed the rise in scores on the new test to inflation allegedly caused by the alleged high stakes. The study

ignored several factors that could have influenced the results, such as differing content, teachers, students, and incentives. Indeed, it ignored most of the factors that could have influenced the results, or speculated that they must have conveniently cancelled each other out, and then declared that high stakes must have done it.

Even nearly two decades later, much of the study remains shrouded in mystery: "The price of admission [to conduct the study] was that we take extraordinary steps to protect the anonymity of the [school] district, so I cannot tell you its name, the state it was in, or even the names of the tests we used." Thus, the study is neither replicable nor falsifiable. An easy solution would be a content match study between the two tests used for comparison. If, as claimed, the two tests represented the same domain (identified, i.e., it could have been [and likely was] as broad as a "grade level" of mathematics from two completely different content frameworks with nonparallel topical sequences), why not support that assertion with some empirical evidence?

The Trouble with Black Boys: And Other Reflections on Race, Equity, and the Future of Public Education
by Pedro A. Noguera, Jossey-Bass, 2008
Reviewed by Joseph A. Hawkins

Pedro A. Noguera is right when he claims that "Black males in American society are in trouble." His latest book, *The Trouble with Black Boys*, provides plenty of well-documented statistical evidence to support how deeply black boys *and* men are in trouble: it is nearly impossible to argue with his data. Nonetheless, the book—beyond the obvious realities associated with how bad things are for black boys—falls surprisingly short when it comes to offering solutions. If things truly are at crisis stage for black boys, then what educators really need—perhaps deserve—is a well-developed book that zeroes in on solutions that work for black boys.[1]

Perhaps *The Trouble with Black Boys* falls short on this front because Noguera, a New York University professor and black man of Latin descent (born in Peru), uses his book to wrestle with too many other issues—Latino youth; immigration; school punishment and its connection to America's prison industry; standards-based reform education; urban schools. The book's subtitle—*And Other Reflections on Race, Equity, and the Future of Public Education*—warns us that Noguera has other intentions. Still, the inclusion of the other issues makes *The Trouble with Black Boys* feel crowded and disjointed.

One upside to *The Trouble with Black Boys* is that Noguera spends considerable time in American schools conducting his own research, and he enjoys sharing it (sometimes by editorializing)—which can be welcome: more researchers in education should reveal how they really feel about what goes on in schools.[2] This context is important, and it certainly provides credibility for Noguera's observations about the ills of schools that have failed to educate black boys and other children of color. Nonetheless, readers might find themselves wishing Noguera provided

more data in some places in *The Trouble with Black Boys*. For example, in chapter 8, Noguera introduces readers to the Minority Student Achievement Network (MSAN), a coalition of suburban and urban school districts founded in 1999 with the intent of pooling resources to research achievement gaps among white, black, and brown students.[3] Using Berkeley (Calif.) High School, an original MSAN member, as an example, he chronicles the school's struggles to serve its disadvantaged black and Latino students. For Berkeley High, Noguera concedes a mix of failures and successes: or, as he puts it, the school offers us a "a glimmer of hope." But if Berkeley High truly is no longer two schools within the same building—no longer "an elite college preparatory school serving affluent White students and an inner-city school serving economically disadvantaged Black and Latino students"—viewing the quantitative evidence would be compelling. Educators and schools struggling with the same problem deserve to hear more of the story, especially an explanation of how such a school narrowed its achievement gaps.

Throughout *The Trouble with Black Boys*, Noguera closely examines parental empowerment and how it impacts poor parents. Noguera firmly believes that poor parents lack power and that attempts to empower them fail in most American schools. Chapter 10 pays brief homage to several workable model programs for parents, including a unique California community partnership that also provides economic benefits. Once again, however, readers may be left feeling as though Noguera could have offered more details.

Readers may also find that *The Trouble with Black Boys* lacks specifics about how to hold meaningful dialogues with people of color. That theme runs consistently throughout the book, and Noguera identifies it as perhaps the main reason that attempts by schools to empower poor students and their parents fail. Others, however, have addressed the issue more successfully. William Ayers and Patricia Ford's collection of essays in *City Kids, City Teachers: Reports from the Front Row* (1996) provides pinpoint guidance. One of the essayists in that book is Deborah Meier, the author of *The Power of Their Ideas: Lessons for America from a Small School in Harlem* (1995), which stands out more than a decade later as a shining example of how to empower both poor students and their parents. Readers who want to learn more about how to hold meaningful conversations about race with people of color can also consult Beverly Daniel Tatum's (1997) "*Why Are All the Black Kids Sitting Together in the Cafeteria?"* or Lisa Delpit's (1995) *Other People's Children: Cultural Conflict in the Classroom*.

From time to time, *The Trouble with Black Boys* surprises—"not all Black males are at risk"—but just when you think Noguera will enlighten us with details about how some black males achieve in school or make it

later in life, he throws a curve. He even speculates in chapter 2 that black boys (and men) who succeed do so simply because of "luck." *Luck!* Yet a fairly well-established line of personal memoirs by successful black men argue against the luck theory. John Edgar Wideman (*Brothers and Keepers*, 1984), Brent Staples (*Parallel Time: Growing Up in Black and White*, 1994), Randall Robinson (*Defending the Spirit: A Black Life in America*, 1998), and Barack Obama (*Dreams from My Father: A Story of Race and Inheritance*, 1995) all demonstrate that for many black men, outcomes depend on more than chance. Both Wideman and Staples do, however, support Noguera's observations that when black males are punished, the punishment is nearly always more severe than what is normally imposed on white males for similar offenses.

In a 1999 book, *African American Males in School and Society: Practices and Policies for Effective Education*, Edmund Gordon wrote, "Some African American males are in trouble, but the African American male condition is not one of universal failure." It is the other side of this coin—*there is hope*—that readers of *The Trouble with Black Boys* may find themselves wishing Noguera had explored in greater detail. Is it unfair to place such a burden on Noguera? Maybe. Maybe not. But given the passion he devotes to the topic of black boys and their futures, perhaps it is fair to expect more from *The Trouble with Black Boys.*

A Personal Note

This past year, for the Bill and Melinda Gates Foundation, I spent months researching the college-preparatory cultures of eleven Washington, D.C., high schools. All the schools in the study are located in D.C.'s poorest neighborhoods—most in the Anacostia section of the city, and all either 100 percent or nearly 100 percent African American. Six of the eleven are public charter schools. (Currently, one of every four school-aged children in D.C. is enrolled in a public charter school.)

I make special mention of the charter schools because it was in a number of these schools that I witnessed hope—perhaps more than a mere glimmer of hope. Hope is a theme that emerges from time to time in *The Trouble with Black Boys*. I witnessed educators listening to students. I witnessed schools putting in place high standards for *all* students—and students striving to meet those standards. I witnessed schools respecting and involving parents, especially as decision-makers. And although I did not find balanced representation by gender (i.e., equal numbers of boys and girls), I nevertheless witnessed the black boys in these schools academically engaged, nearly all graduating from high school and heading off to college.

I share these limited observations not because of some hidden agenda to promote charters over regular public schools; rather, as a black

man and long-time educator and researcher, I believe both kinds of schools play a role in the futures of black children—although I'm not sure that Noguera feels the same. (Oddly, *The Trouble with Black Boys* says little about charter schools; when the term appears in the book, charters are lumped together with vouchers.) But I share my observations because I believe not all is gloomy for black boys: there is hope. There are communities and schools providing black boys with the "chance to be thought of as potentially smart and talented or to demonstrate talents in science, music, or literature."[4] Like Noguera, I've seen that hope—the glimmer of what the future holds or what the present has already achieved. My only wish is that educators would suspend the normal political catchphrases—NCLB is evil, charters cream the talented ones, testing harms—that sometimes blind us to the possibilities.

Notes

1. *The Trouble with Black Boys* does offer solutions—potentially lots of them; however, it fails to pinpoint them as solutions that work for black boys.

2. Some of this research may seem dated, but one could argue that not much in recent times has changed for black boys, and so research findings from, say, ten, fifteen, or even twenty years ago are still relevant.

3. The MSAN website (http://www.msanetwork.org/index.aspx) does not list the Berkeley Unified School District as a current MSAN member.

4. *The Trouble with Black Boys*, p. xxi.

Books Mentioned

Ayers, William, and Patricia Ford, eds. 1996. *City Kids, City Teachers: Reports from the Front Row*. New York: The New Press.

Delpit, Lisa. 1995. *Other People's Children: Cultural Conflict in the Classroom*. New York: The New Press.

Gordon, Edmund W. 1999. Foreword to *African American Males in School and Society: Practices and Policies for Effective Education,* ed. Vernon Polite and James Earl Davis. New York: Teachers College Press.

Meier, Deborah. 1995. *The Power of Their Ideas: Lessons for America from a Small School in Harlem*. Boston: Beacon Press.

Obama, Barack. 1995. *Dreams from My Father: A Story of Race and Inheritance*. New York: Random House.

Robinson, Randall. 1998. *Defending the Spirit: A Black Life in America*. New York: Dutton.

Staples, Brent. 1994. *Parallel Time: Growing Up in Black and White*. New York: HarperCollins.

Tatum, Beverly Daniel. 1997. *"Why Are All the Black Kids Sitting Together in the Cafeteria?" and Other Conversations about Race*. New York: Basic Books.

Wideman, John Edgar. 1984. *Brother and Keepers*. New York: Henry Holt.