education review
a journal of book reviews

Volume 12 Number 2 **January 31, 2009**

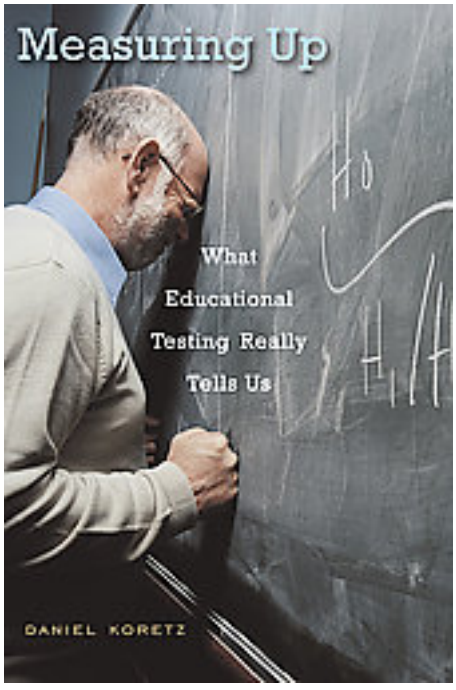
Test Anxiety: An Essay Review of Koretz's *Measuring Up*

Mark Fetler

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, Massachusetts: Harvard University Press. Pp. 353 ISBN 978-0-674-02805-0

Citation: Fetler, Mark. (2009, January 31). Test anxiety: An essay review of Koretz's *Measuring Up*. *Education Review*, *12*(2). Retrieved [date] from http://edrev.asu.edu/essays/v12n2index.html

There seems to be a modest but steady appetite for books about educational testing. A Google search on January 21, 2009 finds 579 English language books published since 1950 with the phrase "educational testing" in the title. A few of these books dish up self-help or scandal, and provide little more than a taste of science and a large serving of mental fast food. Many others are written by scholars for academics and are as learned as they are indigestible. *Measuring Up*'s jacket photo is a biting visual comment on the topic. A college professor, sporting remnants of gray hair, winces as he smacks a dusty, symbol-strewn chalkboard with fist and forehead. He is trying to teach about tests and measurement, but it is a hard sell. Psychometrics is the academic label for "tests and measurement" and it relies on the same kind of objective fact and complex mathematics as the sciences. It is not a diet acceptable to the public or for most college students.



Daniel Koretz

Measuring Up takes a middle course, avoids sensation, but presents technical concepts in ordinary language. The word "really" in the title hints that there is more to testing than what the press sees fit to print. Beyond the topics found in a good introduction, Koretz probes popular views of testing to expose misunderstanding and confusion. Among these is what happens when test scores feed into decisions about rewards and punishments. Students traditionally endure the stress of testing for grades, promotion, graduation, jobs and college admission, and cope, as they must. Given that Public Law 107-110, the No Child Left Behind Act (NCLB), uses tests to evaluate the performance of schools and districts, their teachers, principals, and superintendents now endure similar stress.

Daniel Koretz's education and experience prepared him well to write about testing. He teaches at the Harvard Graduate School of Education. His research focuses on educational testing policy, high-stakes testing, the meaning of score gains, testing students with disabilities, and international testing programs—all of which he explores in this book. During his earlier career, he taught emotionally disturbed students in public schools, earned a PhD in Developmental Psychology from Cornell, worked at the Congressional Budget Office in Human Resources and Community Development, toiled at RAND's Institute on Education and Training, and taught research, measurement, and evaluation at Boston College.

The source of *Measuring Up* is an introductory course on measurement that Koretz teaches. Its goal is to produce well-informed users of tests, not psychometricians. The book contains no mathematical formulas and only a few graphs, charts, and tables. It explains basic technical concepts, such as reliability and validity, and includes more complex topics (e.g., differential item functioning and standard setting methods), but leaves out advanced topics, such as Item Response Theory and its uses. Descriptions of advanced topics would unacceptably increase the length of an already meaty book. The text is reader-friendly in its use of anecdotes, scarcity of citations and absence of the usual pompous bibliography. There are just a few explanatory footnotes and a brief section of end-notes for still-hungry readers.

The conceptual approach does not mean the book is superficial. Koretz goes deep to look at disputes (for example, the meaning of score trends), to review evidence on all sides, and to give his frank opinion. The title of Chapter 1, "If Only It Were So Simple," sounds a theme that repeats throughout. From a distance, testing looks simple. Give students tasks, see how they do, and judge their success. The reality is quite complex.

One of the best lessons from *Measuring Up* is how hard it is to give good tests and usefully report the results. Koretz practices the due diligence of a physician who first describes in grim detail the causes of symptoms, the unappetizing side-effects of treatments, a range of mostly unwelcome outcomes, and then prescribes treatment to a run-down patient. By the time readers turn the last page they may wonder if it is possible to revive this invalid. (Not to fear. The patient has many years to live, whether in good health or poor.) The book's description of the good and the bad, and its mix of technical and policy topics, make it a fascinating and worthwhile introduction to educational testing.

NCLB's Model of Improvement

The title phrase "measuring up" is ordinary language for being accountable. Chief among the disputes that interest Koretz are those that come from NCLB's use of accountability, including the relationship between high-stakes testing and score gains, the abandonment of norms in favor of standards, the inclusion of students with disabilities (SWDs) and English language learners (ELLs) in testing programs, and closing achievement gaps. Review of these topics follows a brief description of testing and accountability under NCLB.

The Merriam-Webster Dictionary gives two definitions for "accountable:" (1) subject to giving an account, answerable, and (2) capable of being accounted for, explainable. To be responsible and to explain are distinct meanings. Unfortunately, in a rush to fix blame or take credit, the value of explanation as a guide for action may be forgotten. Accountability is one way in which resource-laden higher levels of a political system (state and federal agencies) try to control resource-needy lower levels (schools and districts). The idea of accountability is popular, or at least the idea of being unaccountable is unpopular. The reality of accountability, how it actually works, is usually a bone of contention. Because most organizations value independence, tension often results. Ensuing disagreements appeal to reason and research, but usually reflect an underlying power struggle.

There are many types of accountability depending on who is held accountable, by whom, for what, and how. Fiscal accountability is about money. Federal, state, and local governments hold funded schools and districts accountable for spending, whether on qualified staff, well-equipped classrooms, approved textbooks, the availability of rigorous academic standards and aligned tests, or on the type and amount of teaching for certain groups of students. The focus of fiscal accountability is on spending for specific inputs or processes. Examples are audits, program reviews, and accreditation reviews.

Outcomes accountability is about results. How much do students learn? How many graduate from high school, get jobs, go to college, practice life-long learning, are law-abiding citizens, or make positive contributions to society? Perhaps because there are many goals for public schools, and most of them are hard to document and evaluate, accountability programs tend to focus on what is obvious and easily obtained - student achievement test scores. That scores

faithfully reflect learning, and that academic learning predicts all worthy goals is a comforting but risky assumption. A number with a label, e.g., "reading achievement," too easily takes on the reality it claims to measure. A test score is merely a number with a logical relationship to the responses on a test. The score is a better or worse indicator of what a student knows and can do, depending on the quality of the test. Support for inferences about learning depend on understanding the relationship between testing, teaching, facilities, staffing, textbooks, and many other influences on scores.

The 1983 hearings of the National Commission on Excellence in Education and its report, "A Nation at Risk," urged states to increase achievement testing and strengthen graduation requirements. Subsequent federal laws required content standards, aligned achievement tests, achievement standards, and increased testing. The federal approach to accountability before NCLB relied on the publication of test scores, but lacked the teeth thought necessary to bring about change. NCLB added some bite to the federal accountability program.

NCLB requires states to provide for all students in all public schools, challenging content standards that describe at each grade level what students should know and be able to do, and, aligned to those standards, reliable and valid achievement tests given annually in grades 3 through 8 and in high school. States must report results in terms of achievement standards, also aligned to the content standards, that include at least two levels (proficient and advanced) that reflect mastery, and a lower level. Districts and schools must meet state-defined annual targets, that is, make adequate yearly progress (AYP). AYP applies both to all students and to various subgroups. By 2013-14, all students are to achieve at the "proficient" level on reading and mathematics tests. Schools and districts that do not make AYP for two consecutive years are labeled "in need of improvement," and are subject to penalties, such as offering public school choice, supplementary educational services, and eventual takeover.

Underlying NCLB's accountability program is a model of school improvement. Its simple logic persuades some and eludes others. Teaching and testing align with content standards. When the content and methods of teaching correspond to the standards, students have an opportunity to learn what states believe to be important. When a test is based on standards, it yields information about school effectiveness. A schedule for meeting targets and penalties for failure motivates better teaching and learning.

Obtrusive Measures

Koretz's interest in "score inflation" - increases in test scores that do not reflect similar real increases in student achievement - dates from the 1980s. When teachers focus more on the test, their instruction is unlikely to align fully with the standards. Even well aligned tests only cover a sample of material from the standards. Students learn more about the topics on tests and less about other parts of the standards. To explain why this happens Koretz cites Campbell's Law (Campbell, 1975), a tenet of social science research, but less well-known to

educators: "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, 2008, p. 237)

Koretz describes a tell-tale pattern that suggests higher test scores do not reflect real improvements in achievement. Scores on a new test start out low, but show rapid gains over the next several years, eventually leveling out. If a different test is now introduced, the pattern repeats, and over time the scores seesaw up and down. When improvements in teaching and learning are real, and test questions are representative of the curriculum, scores should go up at a reasonable rate and remain high when tests change. The peaks and valleys of test scores show a version of Cannell's (1987) "Lake Wobegon Effect," where all students are above average.

The word "corruption" carries a heavy emotional load, but here has a technical sense that refers to changes in the meaning of test scores when people are under pressure to show improvement. Webb, Campbell, Schwartz, and Sechrest (1966) use a less cutting word, "slippage," to describe what happens. The problem is that tests create as well as measure behavior, attitudes, and abilities. They provoke responses that come from differences in individuals (students, parents, teachers, principals, superintendents) that are irrelevant to the goals of testing (evaluating learning). Webb, et al. (1966, p. 4) note: "... our measures ... tap multiple processes and sources of variance of which we are as yet unaware. At such a stage of development, the theoretical impurity and factorial complexity of every measure are not niceties for pedantic quibbling but are overwhelmingly and centrally relevant in all measurement applications which involve inference and generalization." The way to "reduce slippage between conceptual definition and operational specification," is to use multiple measures, multiple ways of measuring student achievement, some of which are less likely to provoke unintended behavior.

"Corruption" suggests immoral or unethical behavior. By contrast, Webb et al. (1966, p. 16) describe the situation as "role selection." Achievement testing, by singling out individuals (students, teachers, etc.), forces them to decide on a role - "What kind of person should I be ... ?" Their reactions are not so much a matter of defense or dishonesty, but rather a selection from among the many true selves or proper behaviors available. Teachers, parents, principals, etc. believe that their roles include helping students to learn and to perform as well as possible. Their choices may not be among those intended by testing and accountability programs, but it is unfair to assume they are dishonest. If there be fault, it more likely lies with the design of the testing program.

Koretz suggests audit testing as a way to check for score inflation. A low profile audit test that does not have incentives, such as the National Assessment of Educational Progress (NAEP) that produces comparable state level results, may work as a check against state-specific tests that feed accountability programs. The National Assessment Governing Board

(NAGB) reported in 2002 on the use of NAEP to confirm state test results. The report finds that state NAEP results can confirm a general trend of states' own results in grades four and eight reading and mathematics. However, the confirmation is not purely statistical and depends on a "reasonable person" standard. Cautions include differences between NAEP and states in the content coverage of the tests, definitions of student subgroups, demographics, sampling procedures, standard setting procedures, reporting metrics, student motivation, mix of item formats, and test difficulty.

The idea to use NAEP as an audit test is now actual. Regulations issued by the Office of Elementary and Secondary Education (2008a) treat NAEP as a benchmark for comparison with state and local test scores. The regulations require states and districts to include on their report cards the most recent NAEP reading and mathematics results for the state, and to include participation rates for students with disabilities and ELLs. The state report cards must also display the NAEP results for student subgroups.

Glass (2008) and Stake (2007) caution that while NAEP provides the best available test-based evidence of school performance, even that evidence has problems. Originally, NAEP measured knowledge and skills of individual students and summarized the results nationally and regionally. The purposes of NAEP were to further educational research and to verify education reforms coming from the war on poverty and the cold war. Until the late 1980s the program maintained a low profile. In 1988 NAGB began to steer NAEP in the direction of of performance standards, and in 1990 NAEP began to report on individual states. Specifically, fourth and eighth grade reading and mathematics results for individual states and the U.S. are reported biannually as scale scores and percents of students in achievement levels (Basic, Proficient, Advanced). NAEP defers to local preferences for inclusion of ELLs and SWDs, and requirements for local school district participation vary by state, so comparability of state results is open to discussion. As NAEP becomes more intrusive, Campbell's Law comes more into play, and the meaning of the results is more open to discussion.

NAEP scores show an upward trend since the early 1970s, more striking for mathematics than for reading. It is hard to know how much influence Campbell's Law has had on the scores, particularly after 1990 with the publication of achievement level reports for individual states. Variable participation of SWDs and ELLS, as well as voluntary participation of local school districts in some states, make interpretation of the trends more difficult. Displaying results on state and local district report cards calls attention to differences between trends on state tests versus NAEP. However, it also increases NAEP's involvement in "social decision making" and likely contributes to slippage between actual and intended measurement. The idea to use NAEP as a high quality, impartial audit has initial appeal. However, the audit function itself invokes Campbell's Law, and the pursuit of an impartial check is likely to be unending.

There is no perfect solution to the problem of slippage: "Our longing is for data that prove and certify theory, but such is not to be our lot." (Webb, et al., 1966, p. 10) Neither should one completely despair. Any given theory or model, (NCLB's model of school improvement) makes many different predictions. Testing is possible when data are available to check those predictions. The more remote or independent such checks, the more confidence there is in their agreement. Possible sources of school outcome data other than achievement test scores include disciplinary actions, expulsions, grade retentions, remedial education, dropouts, graduation rates, college attendance, college-level remediation, and employment. Using multiple measures for accountability avoids excessive pressure on any one measure.

Koretz (2005) recommends routinely evaluating test results and identifying schools and districts where score gains may be unreasonably large. Patterns in content create opportunities to teach to the test. States can design tests to eliminate patterns in content. Some states refresh large parts of their tests each year. Others design matrix tests, that is, construct multiple equivalent test forms with enough common items to produce comparable student scores. Computer generated tests could produce comparable scores using a unique test tailored to each student.

Evaluations of results and school programs are difficult because every school has a unique history, student body, staff, resources, and surrounding community. What works at one school may not work at another. A work-around strategy is to identify groups of schools that have similar sizes, settings, and demographics. Experts examine the activities and circumstances of successful schools within each group, looking for evidence of successful programs. Research and evaluation would complement NCLB's heavy reliance on test data and formulas.

Likely consequences of implementing Koretz's recommendations are slower and smaller increases in test scores. The dramatic increases seen in "Lake Wobegon" situations take place over a few years. Increases in scores that more faithfully reflect improved teaching and learning probably require longer periods of time and may not fit within AYP timetables. The amount of time needed for improvement is a matter for research and likely depends on many factors, including characteristics of students, parents, teachers, curriculum, schools, communities, amount and use of funding, etc.

Norms Versus Standards

Standards-based reports of test results display labels ("basic," "proficient," "advanced," and the like) to describe levels of student performance. Koretz complains that these reports are more complex than they appear, have a "scientific aura," but are actually quite arbitrary. The labels convey expectations that may be unrealistic. "If educators are to be held accountable for scores, someone has to decide how much improvement is enough. These targets have generally been made up out of whole cloth, with no basis in hard evidence such as normative

data, international comparisons, historical trends showing how rapid improvements are likely to be over time, or evaluations of large-scale interventions." (Koretz, 2008, p. 69)

NCLB's performance standards have three parts: labels, cut-scores, and descriptions. The labels, such as those mentioned above, usually reflect a judgment of the worth of the score. However, states may select any set of labels (for example, "1," "2,", "3"), as long as one of them points to an acceptable level of mastery. Panels of experts identify ranges of test scores (cut-scores) that go with each label. Detailed descriptions of the skills and abilities that embody each performance level flesh out the meaning of the labels. The descriptions may start out as brief "policy" statements provided at the beginning of standard setting and later are refined to match the cut-scores. Or, detailed descriptions may guide the process from the beginning, subject to slight adjustments at the end.

"Arbitrary" in ordinary speech means random or whimsical. Koretz uses the word in an academic sense, meaning not based on hard science. NCLB standard setting is not casual or made-up. The procedures require multiple panels that include teachers who are familiar with students and curriculum. Parents, business-people, professors, and other professionals may take part in order to represent other interested points of view. Where teachers tend to be realistic in setting standards, others may be more ambitious. Standard setting procedures include training, review of performance level descriptions, and examination of test questions. Each panelist judges the range of scores that goes with each performance level. Panelists compare and discuss their judgments and come to a consensus. Before making a final recommendation, they review how the cut-scores distribute the students into the different levels. Often a sub-group of panel leaders or a technical group smoothes the cut-scores for greater consistency across subjects and grades. Finally, policy-makers (State Board, Governor, State Superintendent) review the standards, make adjustments if need be, and approve them.

NCLB's standard setting methods are the same as those used now and in the past to set passing scores on tests for jobs and high school graduation. The methods have been challenged and upheld for decades in courts. One reason for their legal durability is that the job-related uses have clear and specific employment-related criteria for success. Teachers, doctors and lawyers require specific knowledge and skills to practice. Job applicants need certain skills to work effectively. By contrast, the definition of school success is vague. Given the many possible paths through school and after high school (military service, vocational training, college, blue-collar employment, white-collar employment, etc.), it is hard to define what type and amount of learning is adequate. The relevant legal issue for individual students is not "employment-related," rather it is "opportunity to learn" a minimum of knowledge and skills (for example, those needed for graduation). Regarding groups of students and schools, the legal issues are less well defined.

Koretz makes a case for norm-referenced reporting. Where a standard compares a test score to an expectation, a norm compares a score to the actual performance of a group. The norm group may be a nationally representative sample of students. Or, it could be a group of similar schools, a statewide average, a group of states, the U.S., or a group of countries. Commercially available, off-the-shelf, nationally-standardized tests often have norms that provide percentile ranks. For example, a score that goes with the 75th percentile, is higher than three-fourths of the scores in a nationally representative sample of test-takers. The number is descriptive, and does not evaluate performance.

Where some see the descriptive, value-free nature of norms as a plus, for others it is a minus. If the test is easy, an above average score is not necessarily good. If the test is hard, a below average score is not necessarily bad. Moreover, the simplicity of norms is likely more a matter of comfort with old habits than reality. Few people understand the underlying math. Because scores rarely fall into a uniform distribution, equal differences in percentiles mean different things depending on whether they are near the middle, or at the ends of the scale.

NCLB's forerunner, the Title I Evaluation and Reporting System (TIERS), relied on commercially available, norm referenced tests. Publishers used common elements of the curriculum frameworks from many states to plan the content of their tests. Because most states used the same textbooks, it was easy to find similarities. (This idea has ongoing appeal. Even now, NAEP uses a national consensus process to develop the curriculum frameworks that underly its tests.) Publishers administered their tests to a nationally representative sample of students to produce their norms. Bianchini and Loret's (1974) remarkable "Anchor Test Study" of score equating verified comparability of the tests. State and national summaries of the results provided a basis for the TIERS studies.

TIERS and NCLB differ on the comparability of state results and possibility of a national summary. The tests used for TIERS were national in scope and based on a synthesis of curriculum and national norms. Under NCLB each state implements a different set of achievement tests that assess its own content standards. Each state scores its tests using its own achievement standards that are aligned to its content standards. NCLB's approach means that no state's standards, test scores, or proficiency standards are comparable to those of any other state.

Linn (2005) also notes the variability in state's accountability systems. Historically, states assessed different subjects in different grades and attached different rewards and penalties to the results. There has been a general trend to raise the stakes in accountability systems. Some states raised the stakes for individual students or teachers, while others focused on schools. Some of the differences in accountability systems evolved in a patchwork over time, in response to local political needs and conditions that vary from state to state. Linn observes that one goal of NCLB's performance standards is to make it easier for the public to understand reports of test results and to set expectations for acceptable levels of proficiency.

However, the variety of content and performance standards across states makes comparisons hard and a national overview complex.

Special Needs Students

Koretz quotes from a 1997 National Research Council (NRC) report that he helped to write: "The meaningful participation of students with disabilities in large-scale assessments and compliance with the legal rights of individuals with disabilities in some instances require steps that are beyond current knowledge and technology." More sharply, he argues that NCLB's requirement that all students be tested is "unfair to teachers and cruel to students, because it forces them to take tests on which they cannot be successful and to be labeled as failures even if they are working well relative to their capabilities." (Koretz, 2008, p. 308)

These statements reflect the feelings of some parents and teachers, but there are advocates for SWDs and ELLs who disagree. The idea of fairness in education not long ago meant that "all" students - tacitly excluding SWDs and ELLs - should have an equal opportunity to learn. In the past SWDs and ELLs did not participate in regular testing and accountability programs. The educational goals for these students were different, leading to their exclusion from regular classes, sometimes with a defense of hardship on students, parents, and schools. Unintentional effects of exclusion may be less visibility, fewer resources, poor teaching, and less learning. It is clear that NCLB now requires more attention to SWDs and ELLS, and that the situation today differs from 1997.

The "steps" that the NRC report refers to are test accommodations. Accommodations are changes to testing procedures, conditions, or context that do not change the essential meaning of the scores and that help students overcome barriers to participation related to their special needs. Testing accommodations permit the inclusion of those students in accountability programs who might otherwise not be counted. For example, students with limited English language proficiency or blindness, need changes in standardized testing procedures to take most paper and pencil tests. The question is whether the changes cloud the meaning of test scores. How much does reading a reading comprehension test or a mathematics test to an ELL or to a blind student change what the test measures?

ELLs and SWDs constitute a significant portion of public school enrollment. (Hoffman and Sable, 2006) Nationally, in 2003-04 there were 10.6 percent ELLs and 13.6 percent SWDs. While these subgroups are a minority of the total population, they are a large portion of the students targeted by NCLB. The SWD and ELL subgroups intersect and students who belong to both have complex needs and legal protections. Although some ELLs are correctly identified as having cognitive or other disabilities, limited English proficiency (LEP) status is not by itself considered to be a disability.

Students with Disabilities

The Individuals with Disabilities Education Act of 2004 (IDEA) requires SWDs to take statewide tests, with reasonable accommodations, if necessary. (Pullin, 2005) NCLB has a similar requirement and directs states to combine their scores with the scores of all other students and to summarize them separately. During enrollment schools identify and evaluate students who may have disabilities. A team that includes educators and parents creates an individualized educational program (IEP) for each disabled student's instruction and testing that is based on his or her specific needs. Each state sets forth its guidelines for accommodations, provides training, and monitors their use.

Local IEP team decisions on how to test depend on the specific needs of the individual student and the demands of the test. If the student has a severe cognitive disability, for example a neurological defect that significantly hinders normal cognition, the IEP team may decide to provide the student with an alternate assessment that better suits his or her abilities. Alternate assessments align with alternate learning standards. If the student has a moderate cognitive disability that prevents normal progress, the team may decide to provide a modified assessment that aligns with regular grade level standards, but is less difficult than a regular assessment.

Koretz objects to the leeway in identification and classification that IDEA and NCLB allow states. At the extremes, Colorado identifies 9.1 percent of its students versus 15.6 for Rhode Island. There are similar differences in the percentages of students in disability categories. Kentucky classifies 3 percent of students with a "Specific Learning Disability" versus 9.1 percent for Rhode Island. The differences, he believes, are more a matter state and local policy than actual prevalence of disabilities. Koretz may overstate the problem. There are real differences in state populations (health, income, environment) that can influence the prevalence of disabilities. On the other hand, federal limits on funding tend to restrain excessive rates of identification.

Of greater interest to Koretz is the problem that inconsistent identification and classification create for testing. The 1997 NRC report states: "Because disability classifications tell us who may have underlying functional characteristics that are linked to potential score distortions, ambiguities or inconsistencies in classifying students with disabilities have serious implications for assessments ... If classification of a disability is incorrect or imprecise, determining whether the accommodations selected are valid will be difficult." (Koretz, 2008, p. 302)

While the NRC report makes an interesting point, the objection need not be fatal. The disability classification helps in making decisions, but is not definitive. Specific learning and testing needs vary across students with the same disability. Multiple disabilities are not uncommon. Teaching helps to inform testing. Based on the student's specific needs, IEP

teams prescribe individualized supports for instruction that also guide selection of accommodations for classroom, district-wide, and statewide testing.

A sharper criticism is the scarcity of good research on accommodations, alternate assessments and modified assessments. It is not easy to know how effective they really are, whether they actually level the playing field, over-compensate, or under-compensate for disabilities. The number of students with specific accommodations is often small, making it difficult to get statistically definitive results. Accommodations can vary in ways not reflected in a general description. For example, "extended time" may involve more hours, days, or have no limit. Koretz cites a study of the College Board's SAT to suggest that accommodations provide an unfair advantage. However it is unclear that the few disabled students who take the SAT represent the larger group of SWDs, or that the study generalizes to statewide testing programs.

Koretz complains that NCLB has arbitrary caps of one percent for the alternate and two percent for the modified test, limiting the number of students who can opt out of the regular test. Readers of *Measuring Up* could believe that these caps apply to the number of students who may take the tests. In fact, IEP teams are not bound by these caps. The restrictions apply only to the number of students who may count as proficient for the purposes of calculating AYP. There is additional wiggle room in AYP's requirement to test at least 95 percent of students.

Koretz bases much of his criticism on the 1997 NRC report findings. He supports better teaching for SWDs, but opposes NCLB's testing requirements. Some advocates believe that the testing and accountability requirements prompt more attention to SWD's and motivate better teaching, learning, and testing. This belief is evident in the efforts of organizations such as the National Center for Educational Outcomes, (Thurlow, Quenemoen, Altman, and Cuthbert, 2008), and in funding for relevant research. (Office of Elementary and Secondary Education, 2008b) While the problems described in the 1997 NRC report are not solved, the situation now is changed, if not improved.

English Language Learners

Koretz writes, "the issues that arise in testing students with limited English proficiency are in some ways strikingly similar to those we face in assessing students with disabilities." (Koretz, 2008, p. 309) Testing results for English learners may reflect language proficiency, the skills and abilities that the test claims to measure, or both. Actually, NCLB's testing and accountability requirements for ELLs are more complex than Koretz describes.

Two parts of NCLB apply to English Language Learners. Title III requires states to adopt English language proficiency (ELP) standards to guide learning English as a second language, and to administer to all ELLs an English language proficiency test aligned to those

standards. Certain provisions of Title I apply to ELLs identified under Title III. Title I requires academic standards and testing aligned with those standards for all students, including ELLs, with reasonable accommodations if needed, to the extent practicable in the language most likely to yield accurate and reliable scores. Testing results under both Titles of NCLB feed into separate accountability systems that may either reward or punish schools and districts. The test scores may also influence an ELL's grades, promotion, ELL status, SWD status, and funding for services.

When achievement tests directly measure language proficiency, as in tests of reading and/or reading comprehension, accommodations that require reading the test in English or translation clearly change what the test measures, and are not appropriate. Mathematics or science tests might reasonably allow such accommodations. However, some states regard proficiency in academic English as used in classrooms to be an essential part of the test, and accommodations that compensate for poor language proficiency are not appropriate. English only teaching and testing policies may reflect instructional policy, scarcity of resources, or possible bias against non-English speakers.

A difference between SWDs and ELLs is that English proficiency can be taught and learned, in contrast with the permanence of disabling conditions. IDEA, in addition to funding, procedural safeguards, and services, also provides parents and students with legal rights to a free and appropriate education not available to ELLs. A further complication is that some English-speaking students have poor language proficiency skills. To qualify as an ELL under Title III of NCLB, a student's primary language must be other than English. Some students technically do not meet this requirement, but speak regional or ethnic dialects of English. They can talk with their peers but lack the vocabulary and comprehension skills needed for academic classwork or meaningful achievement testing.

Although some ELLs have cognitive or other disabilities, limited English proficiency (LEP) status is not by itself considered to be a disability. If the incidence of disabilities among ELLs is similar to that in the general population, somewhere between nine and twelve percent of ELLs fall under IDEA's rules. For example, in California a state where 27 percent of students are not proficient in English, 9.2 percent of ELLs have disabilities. (Fetler, 2007) The percent disabled ELLs is lower in the elementary grades and consistently rises, with the exception of a grade 9 dip, possibly related to dropout, retention, or inefficient transfer of pupil records to high schools. Understanding the educational needs of these students and the joint requirements of Title III and IDEA with regard to testing is not easy.

While an alternate assessment in the student's primary language may improve meaningful participation in state assessments for some English learners, Abedi (2005, 2007) considers the correct identification of English learners as the most important requirement in providing a fair test. English proficiency test scores are the logical basis for classifying English learners. These proficiency tests measure language skills, not academic achievement. In practice, less

appropriate measures are used, including achievement test scores, immigrant status, number of years in the United States, teacher evaluation, and parent opinion. Differences in the measures result in diverse, often incompatible, definitions of "English learner" across states, districts, and schools.

Simplified English in the directions for administering tests, and in test questions when the topic allows, may be an appropriate accommodation. Depending on state policy and instructional goals, an accommodation for English learners, who are more literate in their primary language than in English, is translation of the test into their primary language. However, translations require extraordinary care in order to compensate for differences in vocabulary, syntax, and culture. If the student receives instruction mostly in English, a better option may be to give the test in that language. Other possible accommodations are access to a glossary, or extra time.

Achievement Gaps

Congress intended NCLB to close achievement gaps between groups of students by focusing on teacher quality, testing, accountability, school choice, and supplementary services, so that "no child is left behind." The 2007 Report of the Commission on No Child Left Behind echoes the thought. "We have a responsibility as a nation to take bold steps to close the achievement gaps that plague our nation’s schools and to ensure that all students are properly prepared for successful and productive lives after high school." (Commission on No Child Left Behind, p. 9)

Koretz looks past the political slogans to ask what it means to close achievement gaps? Average scores for racial/ethnic groups differ, as do those for SWDs versus regular education, ELLs versus English speakers, and poor versus not-poor. The average difference between White and African American groups is growing smaller, but is still significant. The difference between Hispanics and Whites is somewhat smaller, but also significant. Closing the gaps might mean that low scoring groups improve to the level of higher scoring groups or that all groups increase uniformly to a "proficient" level. The goal for AYP is that all students and all groups score at least proficient.

While closing gaps between groups of students is an attractive idea, it appears to require a statistically impossible change in the spread of scores for the entire student population. Group averages tell little about individual students. The spread of scores within most groups is nearly as large as in the population as a whole. Closing gaps requires that the spread of scores, the variance, be greatly reduced. Koretz (2008, p. 139) calls this "the myth of vanishing variance."

NCLB suggests that differences in achievement result mainly from inequities in school quality. In fact there are serious inequities, discrimination past and present, but they only

partly account for the spread of scores. Koretz's analysis of NAEP eighth grade reading and math shows that eliminating differences due to race/ethnicity and poverty (poorly measured by eligibility for school lunch assistance) only slightly reduces the spread of test scores in the entire population. "Simply attributing differences in scores to school quality ... is unrealistic." (Koretz, 2008, p. 142). Beyond NCLB's reach there are stubborn differences in English language proficiency, social class, and general ability that influence test results.

New immigration of ELLs constantly refreshes minority language groups, especially Spanish, but also many others. (Shin and Bruno, 2003) Students without the academic language skills needed to succeed in their classes learn less than their English-only peers. More effective development of English language proficiency and support can help specific cohorts of ELLs to improve their academic learning. However, as long as immigration continues, there will be students with poor English proficiency and ELLs as a group will score lower than native English speakers.

A consistent finding in educational research is that social class (SES) influences student achievement. Traditionally, measures of SES include family income, parent education, and parent occupation. These are in turn linked to personality traits such as achievement motivation and impulse control. Historically, lower SES students score less well than their more privileged peers. (White, 1982) SES correlates with race/ethnicity, ELL, and SWD status. Coleman et al. (1966) found that non-school effects, such as SES, strongly influence achievement. The effects of SES on learning are pervasive and persistent. There are exceptional students, special teachers, and charismatic principals. However, in the main, Coleman's observation stands. Unhappily, U.S. Census Bureau statistics (DeNavas-Walt, Proctor, and Smith, 2008, p. 12) reveal growing poverty. For children under 18 the 2007 poverty rate was 18 percent. Economic programs that promote prosperity for families with school children would almost certainly shrink achievement gaps.

Measures of general mental ability such as intelligence, memory, and abstract thinking are good predictors of individual student achievement and educational attainment. Cronbach (1984, pp. 191 -192) observes that "general ability refers to all-around effectiveness in activities directed by thought. ... Some people perform better than others in solving problems, comprehending events and messages, and learning. They excel then in general ability." As with SES, there are exceptions. People with strong achievement motivation may work hard and over-achieve. Slackers tend to under-perform. However, on the average academic achievement mirrors general ability. Changes in school quality are unlikely to change the spread of test scores related to general ability.

Discussion

NCLB's model of school improvement is simple and plausible. Yet, there are reasons to be wary. Much of Koretz's criticism follows from the effects of high-stakes accountability on

testing - Campbell's Law. One remedy is to use audit tests and research to verify and understand score trends. However, the suggestion that educators intentionally "corrupt" testing programs is a stretch, increases defensiveness, and discourages meaningful research. A more workable approach, suggested by Webb, et al., is that educators choose from among the available actions they believe are legitimate to improve test results. Score inflation may result more from flawed testing programs than from unethical behavior. Matrix test designs, annual item refreshment, and computer adaptive methods would make tests more resistant to accountability pressures.

Educators and elected officials sometimes say that gains in test scores however small are evidence of improvement. There is political value in reporting good news. However, real progress in teaching and learning is better served by an informed view of test results. Koretz's vision seems to focus on work done by independent research and development organizations. He observes that states and schools seldom welcome research done by external organizations. While their work can be valuable when there is good cooperation, there is a case for more broadly based research.

There are enough stories about mechanical failures in testing programs to warrant improved quality control. Overly price-conscious bidding, overburdened testing companies, and stretched state staffs, threaten the quality of testing programs. Schools and districts need training and program audits to discourage improper test preparation and administration. Test scoring and reporting procedures are complex and prone to error. States should require testing contractors to identify and replace questions that are biased or that poorly distinguish between high and low performing students. Periodic alignment studies should verify that the individual test questions and the test as a whole adequately match the standards. Test contractors should undergo comprehensive and regular quality control audits.

After establishing the integrity of the results, studies should examine the significance of changes in the scores. Do the changes meet criteria for statistical significance? From a statistical perspective, are the changes small, medium, or large? Score reports often disregard the sizes of groups of students. A small change may be statistically significant for a large group, but meaningless for a smaller one. Once the statistical meaning of the change is clear, the educational meaning needs investigation. A change of several points may be statistically significant, but might be negligible in the classroom. A slow upward trend may be trivial, or impossible to interpret in the face of demographic shifts and the effects of Campbell's Law.

A reason for the relative lack of research and evaluation of test results may be the scarcity of appropriately trained people. Most large testing programs operate with only a few technical experts who design, implement, and monitor tests, sometimes for many different states. While technology permits large-scale printing, distribution, scoring, and reporting of tests, the expertise is stretched too thin for necessary quality control and evaluation of the results.

NCLB embeds testing deeply into teaching and learning, but does not encourage thoughtful examination of its consequences. Teachers who administer tests, read score reports, and make decisions about instruction and services for students should understand alignment, quality control, and how to interpret test scores. District employees need the technical skills to carry out evaluations and research. State office employees need the skills to check the work of contractors and perform statewide evaluation and research studies.

NCLB funds mainly go into printing, shipping, and scoring and little is invested in people with measurement, research, and evaluation skills. Pre-service and in-service training programs can improve the technical skills of school staff. Federally funded regional education laboratories and research centers can help by providing assistance to states and districts. States and districts should strengthen their own capacity by hiring technical experts in order to write and monitor contracts, to design and to conduct research and evaluations, and to interpret and respond appropriately to the results of testing.

Korezt and Linn complain that states vary widely in their approaches to NCLB testing and accountability. The variety is unavoidable (unless there is a return to a TIERS-like policy) because states with their diverse histories and priorities have primary responsibility for public education. States, not the federal government, provide the bulk of funding for schools. States vary in the funding for each enrolled student, facilities, student-teacher ratios, required hours and days of teaching, and textbooks. A broad view of education would consider the linkage between these elements. Leveling of systems is only likely if testing and accountability actually drive the other components. Staffing, textbooks and facilities are more costly, more politically sensitive, and more likely to preserve differences across states.

A popular view of laws such as NCLB is that they are immutable, engraved in stone. Committees produce a draft, Congress passes it, the President signs, and it is final. More worldly critics understand that NCLB is a work in progress. Large changes come with a reauthorization, but many changes happen along the way. Elected officials and staff respond to special interests (parents, educators, business) to make changes in regulation, guidance, and interpretation. Research programs report on new ways to meet requirements. States' lawyers ask for flexibility and federal lawyers write official letters. Periodically updated guidance (Office of Elementary and Secondary Education, 2009) adjusts technical details. Program review findings continually tweak interpretations. There have been many changes in NCLB since 2001.

What is the future of NCLB? McDonnell (2005) observes that NCLB is a political response to perceptions of achievement gaps. Politicians believe that testing provides information and leverage for holding schools accountable and motivating them to be more responsive to parents and taxpayers. Interest groups and public opinion shape policies at all levels of government. The attitudes of an interest group toward testing heavily depend on perceptions of material benefit. Business groups support testing as a way to increase the supply of trained

workers and to improve productivity. Teacher organizations, focusing on jobs, and civil rights organizations, concerned with student rights, more cautiously support testing. Politicians attend to the public opinion surveys that consistently show strong support for high-stakes testing.

Glass (2008) steps back to look at the larger forces shaping public attitudes. An urban, aging, childless, indebted middle-class wishes to preserve its wealth, consume material goods, reduce taxes, and hence limit the costs of public services, including education. Tests are cheap compared to the expense of facilities, staff, and teaching materials. The low costs of test-based accountability are consistent with continuing political support for NCLB's approach to education reform.

References

Abedi, J. (2005). Issues and consequences for English language learners. In Herman, Joan L. and Haertel, Edward H. (Eds.). *Uses and Misuses of Data for Educational Accountability and Improvement. The 104th Yearbook of the National Society for the Study of Education. Part 2*. Malden, Massachusetts and Oxford, England. Blackwell Publishing.

Abedi, J. (2007). English language proficiency assessment and accountability under NCLB Title III: An overview. In Abedi, (Ed.) *English Language Proficiency Assessment in the Nation: Current Status and Future Practice.* University of California, Davis. Retrieved January 7, 2008 from http://education.ucdavis.edu/research/ELP_Assessment.html

Bianchini, J. and Loret P. (1974). *Anchor Test Study Supplement. Final Report. Volume 31, Project Report.* Berkeley, CA: Educational Testing Service. Retrieved December 15, 2008 from
http://eric.ed.gov:80/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/35/30/b3.pdf

Campbell, D. (1975). "Assessing the impact of planned social change," in G. M. Lyons, ed., *Social Research and Public Policies: The Dartmouth/OECD Conference* (Hanover, NH: Public Affairs Center, Dartmouth College) 35.

Cannell, J. (1987). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average (2nd ed.). Daniels, WV: Friends of Education.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of Educational Opportunity*, Washington, DC: U.S. Government Printing Office.

Cronbach, L. (1984). *Essentials of Psychological Testing*. New York: Harper & Row.

The Commission on No Child Left Behind. (2007). *Beyond NCLB: Fulfilling the Promise to Our Nation's Children.* Washington, DC: The Aspen Institute. Retrieved February 10, 2007 from http://www.aspeninstitute.org

DeNavas-Walt, C., Proctor, B., and Smith, J. (2008). *Income, Poverty, and Health Insurance Coverage in the United States: 2007*. U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau. Washington, D.C. Retrieved December 15, 2008 from http://www.census.gov/prod/2008pubs/p60-235.pdf

Fetler, Mark (2008). Unexpected testing practices affecting English language learners and

students with disabilities under No Child Left Behind. *Practical Assessment Research & Evaluation.* 13(6). http://pareonline.net/pdf/v13n6.pdf

Glass. G. V (2008). *Fertilizers, Pills, and Magnetic Strips: The Fate of Public Education in America*. Charlotte, North Carolina: Information Age Publishing, Inc.

Hoffman, L., and Sable, J. (2006). *Public Elementary and Secondary Students, Staff, Schools, and School Districts: School Year 2003–04* (NCES 2006-307). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved January 7, 2008 from http://nces.ed.gov/pubs2006/2006307.pdf

Individuals with Disabilities Education Act (IDEA). 20 U.S.C. 1400 et seq. Public Law 108-446, (1975/2004).

Koretz, D., (2005). Alignment, high stakes, and the inflation of test scores. In Herman, Joan L. and Haertel, Edward H. (Eds.). *Uses and Misuses of Data for Educational Accountability and Improvement. The 104th Yearbook of the National Society for the Study of Education. Part 2*. Malden, Massachusetts and Oxford, England. Blackwell Publishing.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, Massachusetts. Harvard University Press.

Linn, R., (2005). Issues in the design of accountability systems. In Herman, Joan L. and Haertel, Edward H. (Eds.). *Uses and Misuses of Data for Educational Accountability and Improvement. The 104th Yearbook of the National Society for the Study of Education. Part 2*. Malden, Massachusetts and Oxford, England. Blackwell Publishing.

McDonnell, Lorraine M. (2005). Assessment and accountability from the policymaker's perspective. In Herman, Joan L. and Haertel, Edward H. (Eds.). *Uses and Misuses of Data for Educational Accountability and Improvement. The 104th Yearbook of the National Society for the Study of Education. Part 2*. Malden, Massachusetts and Oxford, England. Blackwell Publishing.

National Assessment Governing Board. (2002). Using the National Assessment of Educational Progress to confirm state test results. Retrieved December 15, 2008 from http://www.nagb.org/publications/color_document.pdf

National Center for Education Statistics. (2007). *National Trends in Reading by Average Scale Scores*. Retrieved April 2, 2007 from http://nces.ed.gov/nationsreportcard/ltt/results2004/nat-reading-scalescore.asp.

National Center for Education Statistics. (2007). *National Trends in Mathematics by Average*

Scale Scores. Retrieved April 2, 2007 from http://nces.ed.gov/nationsreportcard/ltt/results2004/nat-math-scalescore.asp

National Commission on Excellence in Education (1983). *A Nation At Risk*. Washington, DC: U.S. Government Printing Office. Retrieved December 15, 2008 from http://www.ed.gov/pubs/NatAtRisk/index.html

National Research Council, Committee on Goals 2000 and the Inclusion of Students with Disabilities. (1997). *Educating One and All: Students with Disabilities and Standards-Based Reform*. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001 (NCLB). Public Law No. 107-110., 115 Stat. 1425 (2002).

Office of Elementary and Secondary Education. (2008a). Accountability, ssessments, and transparency: How the final Title I regulations support and strengthen the fundamental tenets of NCLB. Retrieved December 15, 2008 from http://www.ed.gov/policy/elsec/reg/proposal/aat.pdf

Office of Elementary and Secondary Education. (2008b). Grants for enhanced assessment instruments. Retrieved December 15, 2008 from http://www.ed.gov/programs/eag/gtepeag.pdf

Office of Elementary and Secondary Education. (2009). Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001. Downloaded January 16, 2009 from http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf

Pullin, D. (2005). When one size does not fit all - The special challenges of accountability testing for students with disabilities. In Herman, Joan L. and Haertel, Edward H. (Eds.). *Uses and Misuses of Data for Educational Accountability and Improvement. The 104th Yearbook of the National Society for the Study of Education. Part 2*. Malden, Massachusetts and Oxford, England. Blackwell Publishing.

Shin, H. and Bruno. R. (2003) Language use and English-speaking ability: Census 2000 Brief. U.S. Census Bureau. Retrieved December 15, 2008 from http://www.census.gov/prod/2003pubs/c2kbr-29.pdf

Stake, R. (2007). NAEP, re*p*ort cards and education: A review essay. Education Review, 10(1). Retrieved December 15, 2008 from http://edrev.asu.edu/essays/v10n1index.html

Thurlow, M., Quenemoen, R. Altman, N. and Cuthbert, M. (2008). Trends in the participation and performance of students with disabilities. (Technical Report 50).

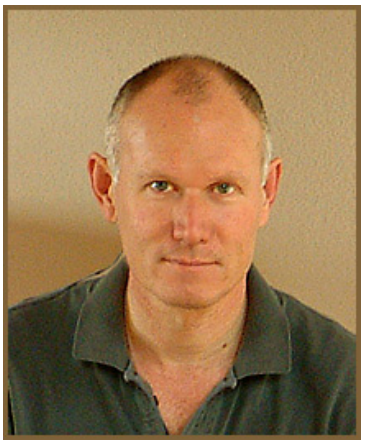
Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Webb, E., Campbell, D. Schwartz, R. and Sechrest, L. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago. Rand McNally College Publishing Company.

About the Reviewer

Mark Fetler

Website: www.markfetler.com
Email: markfetler@gmail.com



Mark Fetler earned a doctorate in Psychology, specializing in quantitative methods, from the University of Colorado, Boulder in 1978. He is retired from California state government where he managed numerous large-scale testing and accountability programs in teacher preparation, higher education, and K-12 public schools. Areas of expertise include policy, technical quality, development, administration, scoring, reporting, and technical assistance. He continues to write, conduct research, and consult on educational testing, and accountability.

http://edrev.asu.edu