Food for Thought: An Essay Review

Mark Fetler

Citation: Fetler, Mark. (2006, October 19). Food for thought: An essay review. *Education Review, 9*(7). Retrieved [date] from http://edrev.asu.edu/essays/v9n7index.html

Herman, Joan L. and Haertel, Edward H. (2005). *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.

Pp. xvi + 383

$39.95 ISBN 1405152605

An unmistakable and controversial change in education in recent years is the dramatic expansion of achievement testing and ambitious school accountability programs in every state. Change of this magnitude naturally draws the notice of the National Society for the Study of Education (NSSE) that has a history of publishing scholarly papers on vital questions in education, such as the following. What are the goals of testing and accountability programs? How do the theory, design, and mechanics of the programs support those goals? How well are the programs working? What improvements are necessary?

Sporting a plain cover and 384 pages crammed with black and white text, the NSSE's 104th Yearbook is a meaty academic volume with answers that are sometimes surprising, but always authoritative. The authors are well known in their fields. Extensive citations enlarge each of the fifteen chapters. Not just a reference tool, the book digs earnestly into the national testing and accountability programs that now dominate school reform and provoke heated debate, more about methods than goals, among educators.

Raising achievement for poor and disadvantaged students, a long-time goal of state and federal government remains a stubborn challenge. The difficulty increases when the goal widens from a student, to a classroom, a school, a district, a state, or the nation. Classroom teachers respond to needs of students and desires of parents, depending on their training, experience, and judgment. Local boards govern schools in response to conversations with parents, teachers,

employers, and community members. State legislatures, under the sway of interest groups that represent parents, teachers, businesses, etc., provide most of the money and write the rules for schools under their authority. Traditions, priorities, resources, and needs differ widely across schools, districts, states and regions. Local and state officials endorse national priorities, but pay closer attention to the tastes of voters and groups that have a say in elections or appointments. On a national level, in the face of overwhelming complexity, what practical methods are there for improving the achievement of poor and disadvantaged students? Accountability is a strategy that has perennial appeal. Although accountability seems simple, the concept has several dimensions.

The Merriam-Webster Dictionary defines "accountability" as "the quality or state of being accountable, especially, an obligation or willingness to accept responsibility for one's actions." Two definitions are given for "accountable:" (1) subject to giving an account, answerable, and (2) capable of being accounted for, explainable. To explain is not necessarily to excuse, but accountability programs that do not make the distinction may sacrifice explanation for the sake of responsibility. Ideally, explanation and responsibility should combine to produce more effective accountability programs than could either one by itself. The need for both responsibility and explanation is a persistent theme throughout the chapters of the yearbook.

Accountability is not new to education and there are many flavors, depending on who is held accountable, by whom, and for what. A traditional and ongoing method looks at the use of money. Providers of money hold those who receive it accountable for how it is spent, whether on able 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 money for specific inputs or processes. Examples are fiscal audits, program reviews, and accreditation reviews. A different method, outcomes focused accountability, looks at results. How much do students learn? How many graduate from high school, get jobs, or go to college?

"Accountability" in the yearbook relates to outcomes. Schools or students are responsible for learning. Consequences come in the form of incentives, often a penalty for low test scores. Accountability programs may or may not have the further goal of explaining the influences on those scores. The "uses and misuses of data" in the title refers to whether the test scores that drive accountability programs are sufficiently reliable and valid for their intended purpose. Most of the chapters address the version of accountability required by the federal No Child Left Behind Act of 2001 (NCLB). However, the yearbook probes beyond NCLB to examine other forms of accountability in classrooms and schools.

NCLB requires states to provide for all students in all public schools, rigorous 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 of achievement (proficient and advanced) that reflect mastery, and a lower level. (The term "standards" in the remainder of this review includes both content and achievement standards.) Districts and schools must meet state-defined annual

targets, make adequate yearly progress (AYP), overall and for specific demographic groups. The state's AYP schedule must show that all students achieve proficiency by 2014. Schools or school districts that consistently do not make AYP face increasingly stiff penalties.

The yearbook also discusses testing for purposes other than NCLB. The federal government's National Assessment of Educational Progress (NAEP) administers achievement tests to a sample of students in each state and nationally in grades 4, 8, and 12. NAEP results for each state provide a point of comparison for statewide testing programs. Many states require students to pass a high school exit test before a district can award a diploma. Benchmark tests are given by school districts several times a year to monitor the effectiveness of instructional programs. Finally, teachers administer classroom tests to monitor individual learning. National and statewide testing programs provide a broad, summative picture of learning overall (national, state, school district, school) or for specific instructional programs or demographic groups, for example, racial/ethnic groups, English learners, or students with disabilities. Classroom tests yield snapshots of each child's progress in a course.

Part one of the book looks at education history and policy as they relate to testing and accountability. The authors discuss ideas about the role of schooling, how politicians translate those ideas into law, and the history of testing and accountability. Part two explores key ingredients of accountability programs that enable those programs to work as intended. Two of these ingredients are the alignment of testing and teaching with standards, and value added measures of learning. The third section explores testing English learners and students with disabilities, and high school exit tests. The book ends with a section that weighs testing and accountability in the classroom.

Part One: History and Policy

Although NCLB put in place an accountability program that looks at test scores, its deeper goal is to encourage equal opportunity for poor and disadvantaged students. Judith Ramaley reflects on the deeper goal, the role of schooling, what the public expects schools to do, and she sets the table for discussion of testing and accountability as strategies for improving education. The public expects schools to improve society and benefit individuals. People benefit through more equal opportunity to get better jobs and to have better lives. Society benefits by promoting common values and helping people to assimilate. How should teaching and testing support these goals? Traditionalists emphasize learning facts, principles, and rules. Progressives focus on creativity, constructing meaning, problem solving, and decision making in the face of uncertainty.

Ramaley favors a progressive style, but she agrees that both approaches have value. Standards that draw from a progressive point of view emphasize critical thinking, higher order cognitive skills, and problem solving. The same standards often include vocabulary and facts that traditionalists favor. Progressives and traditionalists focus on the methods and content of teaching and learning that most directly benefit people. The benefits to society have more to do with goals of equity and efficiency.

Proponents of equity see education as a way to improve the abilities of all students, and

believe that all students should have a common curriculum. Those who favor efficiency would sort students into employment and higher education paths according to their abilities. College bound students should focus on academics. Those students heading to the workplace should learn job-related skills. An unintended effect of an emphasis on efficiency is often an inferior education for those not going to college. While the goal of efficiency remains on the table, the idea of equal opportunity now more strongly influences teaching, testing, and accountability.

Ramaley dwells more on underlying ideas, and less on the politics of translating those ideas into reality. Years ago, educators fought their political battles more over the content and methods of teaching, and less over testing and accountability programs. Today, the growing importance of alignment increases the visibility of testing and draws it more often into debates over education policy. Budgets, perhaps the only sincere statements of policy, reflect the greater importance of testing and accountability in the larger amounts provided for testing contracts, for defraying the costs of giving tests in schools, and for government programs that provide technical assistance to schools. In the present, testing and accountability programs are very much a part of teaching and learning.

The translation of an idea into law is rarely simple, sometimes obscure, and the legislative process is often unappetizing. Lorraine McDonnell describes recent testing policies as a solution to the educational problem of low academic achievement, especially for those who suffer from lack of opportunity and low expectations. The legal solution relies on the belief that testing provides information about achievement for holding schools accountable and motivating schools to be more responsive to parents and taxpayers. High-stakes uses of tests link the results to rewards (more money or recognition for better scores) or penalties (less money, less flexibility, or holding students back from promotion or graduation for poor performance).

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 gain or loss of material benefit. Business groups support testing as a way to improve the efficiency of hiring good workers, and ultimately to improve productivity. Teacher organizations, focusing on jobs, and civil rights organizations, concerned with student rights, more cautiously support testing. Politicians listen to the public opinion surveys that consistently show strong support for high-stakes testing. Majorities of the public support greater school accountability, higher standards, and testing.

Top-down high-stakes tests are attractive because they appear to focus schools on improving achievement. Testing seems to produce quick results that fit officeholders' short-range timetables. Scores on new tests usually go up for the first several years, giving the impression of improvement. Another attraction is the low expense. Tests are cheap in comparison to the high costs of facilities, staff, and materials for teaching. Considering the perceived advantages, politicians are content to stay the course on standards and testing.

McDonnell suggests that informing politicians about the limitations of tests can help to improve accountability programs. Suggestions for improvement should respect underlying political goals, and should be reasonably expedient. For example, statewide or benchmark tests that provide less information about groups of students or educational programs lack the detail and depth of

information useful for making sound promotion or graduation decisions. A defensible high school exit test that lasts hours or days and relies on expensive scoring of essays is not practical for statewide or benchmark testing programs. A test that is good for one purpose is often unsuitable for other uses.

Politicians' appetite for testing and their interest in the quality of the results have grown in parallel over the last several decades. Changes in the intended use of tests motivate greater attention to quality. When tests not only describe achievement, but also play a part in student promotion and graduation, or trigger incentives for schools, then a credible argument is needed for the linkage between testing, teaching and learning. Test results must be more defensible, and their quality receives more attention. Haertel and Herman relate the recent history of testing, with a focus on the tests and standards that NCLB requires. The theory underlying NCLB is that improved learning results from explicit standards for what students should learn, a schedule for meeting learning targets, tests that measure progress toward the targets, and incentives linked to success or failure. The tests measure learning, provide information to guide instruction, and motivate students, teachers, administrators, and parents to work harder. Schools must make progress toward the targets overall, as well as for racial/ethnic groups, poor students, English learners, and the disabled.

A central feature of NCLB's theory of accountability is the alignment of the tests and teaching to standards. When the content and methods of teaching correspond to the standards, students have an opportunity to learn what they need to be successful. When the test is based on the standards, it provides useful information about student progress and the extent to which schools are effective in teaching the standards. Restrictions on the amount of time and money available for giving and scoring tests limit the content of the test to a sample of the material covered by the standards. However, if states build and administer aligned tests, the results should apply to the standards generally.

The effective translation of theory into practice for high-stakes testing and accountability programs requires attention to technical details. Reflecting greater interest in the technical quality of tests, the U.S. Department of Education, Office of Elementary and Secondary Education, (2004) now uses a peer review process to determine whether states meet NCLB requirements and professional standards for tests. The peer review process examines evidence that each state submits in order to show that the tests required for NCLB meet professional standards for fairness, inclusion, alignment, reliability, and validity. The intent is to encourage tests that provide accurate and valid information for holding schools and districts accountable for student achievement against state standards. Independent experts in standards and tests examine the evidence and make judgments about quality. States that do not pass the review face more intrusive federal oversight and loss of funds.

NCLB's accountability requirement increases the pressure on states to improve the achievement of students who perform poorly, but it is not probable that testing alone can accomplish the task. Significant improvement requires more classroom time, better textbooks, and more able teachers. NCLB recognizes these additional needs. However, more time in the classroom, books, and better-prepared teachers are costly additions to any state's education budget. While

politicians and the public agree on the importance of a good education for all students, they may see the expense as equivalent to dining at a gourmet five star restaurant, versus eating cheaper and more popular fast food. Can testing and accountability provide enough leverage for expensive reforms?

Part Two: Design

A simple machine, the lever describes the way accountability systems work. The parts of the machine are a bar with a moveable object at one end, a mover at the other end, and a pivot. The lever multiplies the mover's force depending on the closeness of the pivot to the object. Moving a heavy object a short distance requires moving the other end of the lever a longer distance. For the machine to work it is essential to arrange the parts of the system as described. The same is true for accountability systems, except that the arrangement of the parts, their alignment, is more complex. Robert Linn describes those features of alignment that make accountability systems work.

There are several ways to evaluate alignment. A customary approach is to match the content of test questions with the broad content categories in the standards. For example, a mathematics item might match with "number sense," or with "problem solving." States' standards vary in their level of detail, but those with more specific descriptions of content provide better direction for the development of tests. Looking at broad content categories is a step in the right direction, but it is not enough to guide teaching and test development. A more complete evaluation of alignment looks at additional facets of content, including: complexity, ranging from recall of facts to conceptual understanding; the relative emphasis and range of topics; and linguistic features, the characteristics of the language in the standards.

The federal peer review of state's standards and assessment systems requires states to provide evidence that their tests meet the following conditions for alignment: cover the full range of content in the standards; measure both the content (what students know) and the process (what students can do) aspects of the standards; reflect the same degree and pattern of emphasis apparent in the standards; and, reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the standards.

Linn observes that the differences in states' standards, tests, and accountability programs make comparisons difficult. Historically and legally, education is more the responsibility of state than federal government. States and not the federal government provide most of the money for education and earmark it for local priorities. Although differences in state standards and testing programs are real, national tests can assist in comparing state programs.

NAEP administers fourth and eighth grade tests in reading and mathematics every other year in all states. These assessments, in combination with surveys of student, teacher, and school background characteristics, yield a wide array of state-by-state findings. However, NAEP's biannual testing schedule and sparse sampling of schools and students restrict the usefulness of the results. Some states object to comparisons because their standards differ from NAEP's national consensus standards. A reply to this objection is that NAEP and state reading and mathematics tests likely measure the same constructs. The similarity of the constructs is an imperfect but acceptable basis for

making state comparisons.

Assuming that the goals of accountability are to motivate and improve learning, it is useful to understand the reasons why test scores change. The ability to give an explanation depends on the design of the testing program. An illustration of one such popular design follows. A state arranges to administer different tests for each required grade. For example, one year all third grade students take the third grade reading test, resulting in an average score for the third grade. Next year, those students promote and take the fourth grade test, producing an average score for the fourth grade. A new group of third graders takes the third grade test, and produces a new average third grade score for comparison with last year's average score. Even though the third and fourth grade reading scores look similar, the tests are different, are not comparable, and do not measure growth across grades. This common design allows an estimate of the status of a school's grade in any given year and changes in status over years. It is possible to look at score trends within a grade across years, but not across both grades and years.

A change in the average third grade score may reflect a change in learning, or it may reflect a changing population with different numbers of poor students or English learners. Each year teachers retire from service and new recruits enter the classroom. Communities grow or shrink. Businesses hire more people or shut down. Economies spread prosperity or poverty. The context in which schools operate changes constantly. Year to year changes in average test scores may reasonably result from changes in the mix of students, the context of schooling, or from teaching. Long-term economic or demographic trends may influence test scores over years. Testing systems that measure status do not yield estimates of growth that take into account changing student or school characteristics. The chapter by Choi et al. describes methods for estimating growth that do take into account the characteristics of students and schools.

Choi's methods involve tracking of test results for individual students from year to year. In order for tracking to work, it must be possible to compare test scores meaningfully across years. For example, if the third and the fourth grade reading scores measure similar skills and are on a common scale, it is possible to estimate a student's growth from the third to the fourth grade. The tests and the common scale that spans several grades must align with content standards that hang together logically across grades. If the tracking includes information about classrooms and schools, it is possible to calculate not only individual growth, but also average classroom and school gains. Testing designs that allow these calculations of growth along with statistical adjustments for student or school background characteristics are called "value added models."

Consistent with the design of many statewide testing systems, NCLB's basic approach to accountability depends on estimates of school status in any given year and improvements in status over years. Choi suggests that research on student background factors and context is needed to interpret the results of status-based systems. The U.S. Department of Education does allow value added models in states that have student score tracking programs.

An area of research for value added models is the ability of a common scale to measure change across more than a few grades. Teaching goals, methods, and materials change across grades to keep pace with maturation and prior learning. The tests and standards must reflect these changes.

After several years it is no longer meaningful to compare scores across grades, for example, comparing first grade reading performance in phonemics with comprehension of a story's main idea in the sixth grade. High school algebra and geometry do not easily compare to arithmetic taught in elementary grades. Measures of growth in a subject appear to have the most meaning across adjacent grades, but become harder to understand with greater separation.

Another area of research relates to the appropriateness of statistical adjustments. Leveling the playing field with regard to poverty recognizes and compensates for the greater barriers to learning for poor students, and the resources that schools must provide to overcome those barriers. However, making those adjustments suggests that schools will not be held accountable for improving the learning of poor students, but only for meeting a lower, adjusted standard.

For accountability to work it must be true that higher test scores reflect better teaching and learning. Support for this belief comes from alignment of the test with standards and teaching. More support comes from student tracking, common scales, standards that progress logically across grades, and statistical adjustments to reduce the influence of changes in demographics. These supports do not guarantee that scores are meaningful. Teaching to the test is inevitable when accountability links scores to incentives, whether greater visibility, rewards, or penalties. Koretz describes a threat that higher test scores could reflect coaching or teaching to the test and not real improvement.

Koretz describes a tattletale pattern that suggests teaching to the test. 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 teachers devote more time to material 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. The seesaw pattern is not seen on external tests, such as NAEP, that lack incentives. When improvements in teaching and learning are real, scores go up at a reasonable rate and remain high when tests change. How can states promote teaching that aligns with standards in the face of an incentive to focus on the specific material in tests?

Koretz recommends evaluating test results and identifying cases of severe score inflation. For example, a state can examine the gains made by schools in order to identify those that are unreasonably large and then investigate. States can also design tests to eliminate patterns in content or weighting of topics over time. Eliminating such patterns removes a temptation to teach to the test, but has additional costs for developing tests and maintaining the comparability of scores. Using multiple measures for accountability avoids excessive pressure on any one measure. Finally, expert judgment can improve accountability programs that now depend on simple formulas to make decisions.

Evaluation of results and expert review of 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, poverty levels, or percentages of English learners. Experts examine the programs of successful schools within each group, looking for evidence of successful programs. Evaluations can benefit from careful blending of subjective and objective research methods. Analysis of test scores requires reliable and valid data and quantitative methods experts. Evaluating educational programs requires experts trained in qualitative methods who can consistently make unbiased judgments. Together, objective and subjective methods encourage a balanced research perspective.

It is tempting, but wrong, to assume that a test score is a precise measure of learning, whether for individuals or groups. The statistic that forty percent of the students at a school are proficient in reading looks like a simple and pure measure of performance. However, statisticians tell us that test scores, like the truth, are rarely simple and never pure. Every score contains a measure of error that stems from the conditions of testing. Combining scores compounds the errors. David Rogosa writes about the statistical properties of the scores that NCLB accountability programs use. He presents sketches of situations involving measurement and judgment, where users of test results overstep the bounds of statistical certainty. Creating NCLB accountability measures requires summarizing test scores for a school, for example, calculating the percent of students who are proficient in reading or computing weighted combinations of scores. Judging involves making a decision, depending on a measure, whether a school met criteria for success. The challenge is to use measures to make judgments in a way that is statistically sound. While there are statistical rules of thumb for simple situations, measures for school accountability are more complex, and call for the professional judgment of statisticians.

Part Three: Fairness and Consequences

The idea of fairness in education is akin to the idea of equal opportunity found in NCLB's requirement that standards and testing apply to all students. All students have a seat at the table. In the past the goal of inclusion did not apply to students with disabilities or English learners. The goals and teaching for these students were different, leading to exclusion from testing and accountability programs, sometimes with a defense of hardship on students, parents, and schools. Unintentional effects of exclusion are less visibility, fewer resources, poor teaching, and lower achievement. While inclusion in testing programs is one facet of equal opportunity, it raises logical questions about fairness in testing. Some students have special needs, for example, limited English language proficiency, brain injury, or blindness that hamper meaningful testing. For these students meaningful participation requires changes in the usual standardized conditions of testing. How much can schools alter the conditions of testing, in order to include more students, before the scores no longer usefully measure learning? What amount of English proficiency does a student reasonably need in order to take tests meaningfully in English? Jamal Abedi and Diana Pullin discuss these questions as they write about the inclusion of English learners and students with disabilities in accountability programs.

NCLB requires testing limited English proficient students with reasonable accommodations,

to the extent practicable, in the language most likely to yield accurate and reliable scores. Correct identification of English learners, according to Abedi, is 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. The use of these other measures to make classification decisions varies widely within and across states. Differences in the measures result in diverse, possibly unsound, often incompatible, definitions of "English learner" across states, districts, and schools.

Accommodations are changes to test procedures, conditions, or context that do not change the essential meaning of the scores and that help students overcome barriers to testing related to their special needs. All tests that are given in English measure not only performance on a standard, but also English language skills. The results for English learners may reflect their poor English and not the skills and abilities that the test claims to measure. The same is true for native speakers of English who have poor language skills. This threat becomes more serious as the language demands of the test increase and as levels of English proficiency go down. Accordingly, using simplified English for testing is a logical accommodation. Test results are more accurate to the extent that the language of the test can be simplified, not only for English learners, but also for all students. 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. Translations require care in order to compensate for differences in vocabulary, syntax, and cultural context. A better option is to give the test in the language of instruction. Other accommodations that make sense for English learners are access to a glossary, or extra time.

The question for students with disabilities is not whether, but rather how, they take part in testing for NCLB accountability. The federal Individuals with Disabilities Act of 2004 (IDEA) requires students with disabilities to take statewide tests, with reasonable accommodations, if necessary. NCLB has a similar requirement and directs states to combine their scores with the scores of all other students and to summarize them separately. A team that includes educators and parents creates an individualized educational program (IEP) for each student's instruction and testing, based on their specific needs. Each state sets forth general procedures for accommodations, and the local IEP team makes decisions for individuals. If the student has a severe cognitive disability, for example a neurological defect that severely hinders the ability to think, the IEP team may decide to provide the student with an alternate assessment that better suits his or her abilities. Alternate assessments must align with standards, provide valid and reliable results, and be included in the accountability calculations. Knowledge about the tests and research on accommodations should inform decisions about accommodations and alternate assessments. In practice, IEP teams decisions often rely on imperfect knowledge of research, intuition, and sympathy for the disabled student.

Pullin notes the meager research on accommodations, alternate assessments, and accountability for students with disabilities. The number of students with specific accommodations is often small, making it difficult to get statistically sound 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. Definitions of individual needs differ, depending on the specific disability and on the technical skills available to the IEP team. For example, the "learning disability" category covers a large number of different problems that are not clearly distinguished. In the same vein, "English learners" have varying degrees of English language proficiency, different language and cultural backgrounds, and differing amounts of literacy in their primary language.

An alternative to the use of accommodations is to provide "universally designed" tests that have fewer barriers to participation, for example, removing time limits and using simple English whenever possible. Standardized tests try to control as many of the conditions of administration as possible, in order to provide the same testing experience to all students. This rigid approach improves the reliability of test scores, but excludes students with special needs. Universal design works with more flexible conditions of administration to include more students. A problem for universal design and the use of accommodations is to balance greater inclusion against greater difficulty in interpreting the meaning of test scores.

Although NCLB uses tests to hold schools accountable, many states hold individual students accountable by requiring that they pass a standards-aligned exit test to promote to the next grade or graduate. In the past, states advertised these tests as setting a high bar, recognizing and rewarding achievement, and selecting qualified students for college. Now, states pitch them as a way to improve the achievement of all students. John Bishop describes studies that evaluate the use of exit tests in different countries. In the U.S. only students who pass can receive a diploma, the tests measure a relatively narrow range of achievement, are pass/fail, and the passing standards are set low enough to allow almost everyone to pass after several tries. Exit exams in Europe measure a full range of achievement, include more difficult questions, relate to specific courses, are not pass/fail, and the results signal different levels of achievement, influencing grades, jobs, and college admissions.

European tests correlate with higher achievement in mathematics and science, smaller gaps between high and low socioeconomic groups, higher minimum standards for entry into teaching, higher teacher salaries, and more teacher specialization. European tests increase achievement without decreasing enrollment or graduation rates. U.S. tests have a much smaller effect, if any, on achievement and reduce the number of students getting a diploma. However, students in states with exit tests tend to earn more after graduation than students in other states, perhaps reflecting employers' improved perceptions of graduates. Bishop examines the reasons for the success of European tests. He suggests that the wider range of measured achievement acts as an incentive to all students. By contrast, the lower, narrow range of achievement in the U.S. tests focuses attention on a smaller group of low-achieving students. European tests do not prevent graduation, but rather document a level of achievement for consideration by employers and colleges. European end-of-course tests show how individual teachers succeed with their classes. The U.S. tests cover material taught in many classes, so that individual teachers are not accountable for their students. Finally, European tests are more challenging and provide a more worthy goal for teachers and students.

Melissa Roderick, Jenny Nagaoka, and Elaine Allensworth describe Chicago's use of tests to

promote or retain students. In Chicago's program students in grades 3, 6, and 8 face summer school and retention if they do not get minimum scores (marking the bottom one-fifth of students nationally) on tests in reading and mathematics. Roderick reports no change in how teachers teach and engage students, or in the development of their professional skills. The testing requirement stimulates more teacher and parent social support for low achieving students, but not higher expectations. Summer school motivates short-term test score gains in the sixth and eighth grades. However, retention does not improve the achievement of third graders, and correlates with lower sixth grade achievement. The Chicago experiment displays a troubling side effect. Retained students seem to end up in special education, where the testing requirement no longer applies.

Testing for high school graduation or grade promotion in the U.S. demonstrates the law of unintended consequences. The public expects the tests to protect the value of a diploma and to discourage social promotion. Instead of providing a respectable floor, high school exit tests establish a low ceiling for achievement. Teaching to the test inflates scores. The end of social promotion increases the number of students placed in special education, where high expectations no longer apply. Bishop and Roderick do not criticize the goal of high expectations. However, their studies show the difficulty of maintaining ambitious goals when there are harsh consequences for failure. Both authors suggest that part of the solution is to improve teacher training and classroom practice.

The chapter on teaching and teacher quality by Linda Darling-Hammond and Elle Rustique-Forrester enlarges the point made by Bishop and Roderick et al. Unforeseen negative consequences can result from accountability systems that use few measures and do not attend to improving teaching and teacher quality. States and districts that combine aligned tests with better teacher training raise student achievement on multiple measures, even without strong incentives. When states and districts do not improve teaching, the achievement of low performing schools languishes. Areas of teacher training that produce positive results include: treating testing as a core part of teaching and teacher training; combining large-scale external tests with classroom tests; using more detailed tests to get at the root of teaching and learning problems; and making training in the standards a requirement for teacher licensing, certification, and ongoing evaluation.

Part Four: Better Practice

The alignment of standards, teaching, and testing is meant to act like an enlivening tonic throughout a state's school system. For the tonic to work, teachers must practice alignment daily in their instruction and testing. The chapters by Heritage and Yeagley and by Shaw look at practice in the classroom. External, large-scale tests, given annually, and district benchmark tests, given several times a year, produce broad measures of learning. Classroom tests yield more detailed, timely information about a student's progress. Ideally, the standards guide statewide, benchmark, and classroom tests. Unfortunately, many teachers lack training in the standards. Even larger gaps in training exist in developing aligned tests, reporting the results, and understanding the meaning of the scores.

Accountability programs can use either structural or functional methods. Structural methods for improving schools, thought to be more hard-nosed and objective, focus on distinct parts of a system (standards, teaching, tests, flow of information, incentives) and their alignment. By contrast, functional methods, seen as softer and more subjective, focus on the people in schools (students, teachers, managers), their roles (learning, teaching, coordination), and how they interact to accomplish their goals.

Eva Baker presents a functional alternative to NCLB's more structural approach. Her proposal focuses less on content standards and more on skills and knowledge. It focuses less on the types of test questions (multiple choice, open-ended), and more on types of thinking skills. It describes performance less in terms of a score or category and more in terms of the kinds of skills and abilities that correspond to degrees of proficiency, and the ability to transfer the use of skills to new situations.

In Baker's system learning is central. She describes five groups of learning tasks: problem solving, content understanding, communication, teamwork, and metacognition. There are many ways to test the tasks in each of these groups. A math problem might focus on problem identification and use an open-ended question. Or, it might involve several steps that end with selecting a correct answer. Research on each group of tasks helps to specify what thinking skills and strategies to use for testing a particular topic. These thinking skills can connect different tests, whether teacher-made classroom tests, district benchmark tests, or external statewide tests. Ideally, information from all of these tests is available to teachers, parents, students, and others to help with teaching, learning, evaluation or accountability.

Baker recommends working with a small number of standards, fewer than usual for a grade or subject. Most state content standards are ambitious and broad, resulting in gaps between official goals and what actually appears on a test. She suggests several ways to reduce the gap: limiting the number and type of standards to the most significant testable ones; using articulated frameworks to encourage common expectations across test producers; providing detailed descriptions of the content to be tested; and permitting teacher given tests to count for the purpose of accountability. She opposes broader testing that covers the entire range of goals found in most standards documents.

Teachers use classroom tests for grades, monitoring student progress, and diagnosing strengths and weaknesses. Questions arise if accountability programs are to use teacher-made tests. Should student grades be consistent with the results of external statewide tests? How can states assure the quality of teacher-made tests in respect to their content, and scoring? How can states cope with conflicts of interest if the results benefit or penalize teachers? Baker and other authors in this book call for more research on the effectiveness of accountability systems. The role of teacher-made tests in accountability is one such promising area. Teacher-made tests are a strong, direct influence on learning because they are more timely and relevant to student needs than district benchmark tests or statewide external tests.

Digression

Herman and Haertel end their preface with the dry understatement that the book describes the "range of factors" underlying accountability. Well beyond a "range of factors," the authors pour out oceans of detail and map vast expanses of territory. However, a few basic themes repeat across chapters. Alignment, the correct arrangement of standards, tests and teaching, provides the leverage for accountability to work. The language demands of standards and tests are an aspect of alignment needing further attention. States and school districts need technical expertise in testing and measurement, in order to evaluate results and to use value added models. Teachers need training in standards and testing to improve student learning. Accommodations and universal design of tests, for inclusion of all students in accountability, are a change from past methods of standardized testing and need study. Finally, states vary in their standards, tests, and accountability programs. What is the significance of this diversity?

Alignment. The idea behind alignment is that matching the content of the test with standards provides leverage to improve teaching and learning. To focus attention on the standards, alignment must consider broad content areas, as well as the complexity, emphasis, and range of topics. Abedi and Linn mention language demands as one further aspect of alignment. Language demands cover the linguistic features of sounds, syllables, words, phrases, and grammar, as well as particular uses of language, for example, identification, labeling, definition, comparisons, persuasion, or evaluation. Language demands are relevant not only for English learners and some students with disabilities, but also for the larger group of disadvantaged students and others who have poor English language proficiency. Some disadvantaged students from English speaking homes rival English learners in their lack of language skills. Better learning for all students depends on improving mastery of language and communication skills. Traditional tests of reading and writing do not provide useful test scores in the low range English proficiency shown by many of these students. These tests often provide no information on the listening and speaking skills that are necessary for successful communication and that support reading and writing.

There is a history of research on the content validity and alignment of tests that are built by matching the questions to the standards. However, there seems to be less research on the alignment of teaching to the standards. For example, teachers may have a vague sense that a test aligns with the content in the standards. However, they likely do not understand that alignment requires a match on the dimensions of breadth, complexity, and linguistic features of the content to be taught and learned. A topic for further research is the extent to which teachers' understanding of alignment actually changes practice in the classroom in desirable ways and improves learning of the standards.

Evaluation of Testing. States often praise any gain in test scores, however small, as evidence of significant improvement. There is political value in reporting good news, and such reports may help to motivate parents, teachers, and students. However, real progress in teaching and learning likely depends on a deeper understanding of the results. Several authors comment on the need for evaluations of the results of testing.

Schools and districts need evaluations to discourage improper test preparation, administration, and score inflation. Moreover, procedures for test scoring and reporting 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. Alignment studies should verify that the individual test questions and the test as a whole adequately fit the standards. If the tests are to be comparable across years, testing contractors should design sound equating plans, monitor implementation, and analyze the results. Evaluations provide necessary quality control checks on the work done by test contractors and states.

After establishing the quality of the results, further studies help to explain the meaning of changes in the scores. A next step is to check whether the changes meet criteria for statistical significance. Do the changes fall within a range obtainable by chance? 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 is a topic for investigation. A change of several points may be statistically significant, but might be negligible in the classroom. A steady upward trend may be trivial if it takes too long to reach the goal. What are the causes of score trends? Compared to teaching, how large are the influences of the community or the economy?

Accommodations and Universal Design. Accommodations remove barriers to participation in standardized tests for students with disabilities and English learners. Examples are the use of large print, extended time, or assistive devices, that relate to students' specific needs. Any change in the standard procedure for giving the test increases the statistical uncertainty of scores and affects their meaning. Changes beyond a certain point make the scores useless. The task of research, in combination with expert judgment, is to draw the line between accommodations that provide reasonable access and those that go too far.

An alternative to research on accommodations is to create tests exhibiting universal design principles that lack the barriers found in standardized tests. It is easier to include more students if the test has no time limits, is administered on a computer screen with variable sized fonts, can be given aloud, and uses simplified English. Designing and developing universally designed tests requires careful thought about the measurement of constructs and the extent to which the design features might interfere. Adopting universal design principles may require rethinking the standards that inform tests and teaching. The role of language demands in the communication of knowledge needs more attention. For example, given that blind students "read" books using the spoken word, why should reading standards forbid spoken versions of reading comprehension tests? Or, does less demanding language in standards and tests produce better measures of mathematics and science?

Training. External testing programs might operate with a few technical experts who design, implement, and monitor tests. One effect of NCLB is to embed testing deeply into teaching and learning. Checking alignment, evaluating results, and conducting research on accommodations require people with specialized skills and training. 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.

Under NCLB the federal government and states are paying for more testing and accountability, but it is not clear that more people with the appropriate skills are available to work on those programs. 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. Moreover, 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.

State Diversity. Several authors comment that states vary widely in their approaches to accountability. Some standards have more detail than others, or progress more or less logically across grades. Testing formats emphasize multiple choice questions, or open-ended response tasks. Test scores provide status information, or value added results. The variety in testing spills over into state accountability programs. The diversity is unavoidable because states, not the federal government, have primary responsibility for public education. A coping strategy is to evaluate state programs and share the results with the public. The federal standards and assessment peer review process moves in this direction. Another option is for interested states to meet periodically and collaborate. The Council of Chief State School Officers runs state collaborative meetings that produce useful dialogue and resources on technical issues, accountability, testing, English learners, students with disabilities, and other topics. NAEP can model sound testing practices for states, and more information for making comparisons could come from expanded NAEP testing.

Conclusion

The yearbook points from a crisply described past towards a hazy future. It is easier to look back than to peer ahead. The editors agree that the book not only describes the present state of the art, but also shows that we can do better. Leading up to the reauthorization of NCLB in 2007, there should be a discussion of what works well and what needs changing. Participants should include parent, teacher, administrator, and school board organizations, and should also include education research and evaluation groups. The distinguished authors of this yearbook provide an extensive and useful stockpile of information, food for thought, for a discussion of amendments to NCLB.

References

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

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

Office of Elementary and Secondary Education. (2004). Standards and Assessments Peer Review

Guidance: Information and Examples for Meeting the Requirements of the No Child Left Behind Act of 2001. Washington D.C., United States Department of Education.

About the Reviewer

Mark Fetler
markfetler@gmail.com

Mark Fetler earned a PhD in Psychology from the University of Colorado in 1978. He is retired from employment in California state government where he worked on programs in teacher preparation, higher education, and K-12 public schools. His interests include educational research, assessment, and accountability.

Editors: Gene V Glass, Kate Corby, Gustavo Fischman