education review
a journal of book reviews

Volume 10 Number 9 July 10, 2007

NCLB Blue: An Essay Review

Mark Fetler

The Commission on No Child Left Behind. (2007). *Beyond NCLB: Fulfilling the promise to our nation's children*. Washington, DC: The Aspen Institute.

Pp. 230 $16.85 (Shipping & handling) ISBN: 0-89843-467-X

Citation: Fetler, Mark. (2007). NCLB blue: An essay review. *Education Review, 10*(9). Retrieved [date] from http://edrev.asu.edu/essays/v10n9index.html.

Introduction

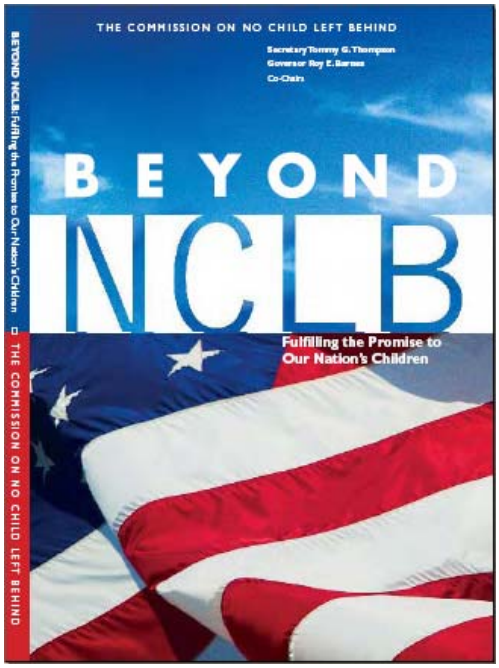
Ceremonies in government often accompany notable events and invest them with meanings that are more profound than the bare facts support. Such ceremonies demand, and receive, a degree of public support, respect, and gravity. The ritual trappings that adorn legal hearings, roll calls, and ballots overlay and dignify arcane voting procedures for elections or legislation. Likewise, the venerable blue ribbon panel, with its magisterial board, solemn hearings, and weighty report, enacts a ceremony that both polls public sentiment, and cloaks the legislative process with a decorum it might not otherwise display.

The phrase "blue ribbon" connotes judges who are deemed to be exceptionally qualified to weigh evidence on their assigned topic. While their public résumés are gilt-edged and adorn the report, the vetting of their political views and the processes of nomination and appointment to seats on the panel are discrete affairs. The panelists consume vast amounts of information at open forums and closed executive sessions. In a process that is often at best translucent, they sift, mix, and knead the information into a report. After the ink dries, and the

insight of the appointees has been lauded, lawmakers may adopt some or none of the recommendations. Whether the factual weight of the report visibly sways elected officials, the political heft of the panel confers an aura of authority on the result.

The need for a blue ribbon panel grows in proportion to the controversy and scope of the law in question. Few laws in education are more controversial and have wider impact than the No Child Left Behind Act of 2002 (NCLB) that dramatically expanded achievement testing and accountability programs across the United States, leaving many states and schools to struggle with its requirements. Thus it is ordained that a national blue ribbon panel is attending to the looming sunset and reauthorization of NCLB.

The Commission



The 2007 report of The Commission on No Child Left Behind, "Beyond NCLB: Fulfilling the Promise to Our Nation's Children," suggests changes in Public Law No. 107-110 of 2002, an act of Congress intended 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 politically connected Aspen Institute, whose motto is, "timeless values, enlightened leadership," provided an administrative home for the Commission and funneled resources from various prominent, national, non-profit foundations to support its activities. Tommy Thompson, former republican governor of Wisconsin, and Roy Barnes, former democratic governor of Georgia co-chaired the Commission. Eminences from business, higher education, government, and K - 12 education populated the panel. The fifteen members included three women, three African-Americans, and two Hispanics. Public hearings, roundtable discussions, site visits, email, and literature reviews provided the raw data that were sorted, selected, and massaged into its final report.

The tone of the Commission's report echoes the martial and heroic resonance of the slogan, "no child left behind." The metaphor suggests a country engaged in a battle, fighting with determination to reclaim children who are held hostage. The war-like theme appeared earlier in the 1983 blue ribbon commission report, "A Nation at Risk," and since then has remained a leitmotiv of federal policy. Whether or not the military model is a good fit for public schools, similar themes, clad in appropriately dramatic rhetoric, appear early in the Commission's report. For example, the Forward begins with the following thought.

> "We cannot afford to sit idly by and hope that things will improve. 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. Failing to take sustained action will not only result in the continued tragedy of unfulfilled potential, but will also threaten our nation's economy and future competitiveness in the world." (NCLB Commission Report, p. 9)

The NCLB slogan suggests that the war's goal is to save children, and the legislation specifically states that it is to "close achievement gaps." Various players fill the roles of combatants and weapons in this epic struggle. The protean enemy seems to be low achievement, or more logically, poverty or discrimination, but at times could also be teachers, administrators, or public schools. The table of contents helps to flesh out the metaphor with chapters on teacher qualifications, accountability, supplementary student support, school choice, and assessment. Additional chapters discuss the needs of English language learners, students with disabilities, and migrants.

This review focuses only on the Commission's findings on assessment and accountability, topics that dominate center stage in the implementation of NCLB, the Commission's report, and debates about the effectiveness of the law. In 2005 Herman and Haertel edited, and the National Society for the Study of Education (NSSE) published, a landmark volume of scholarly papers by prominent researchers on the use and misuse of test data for NCLB accountability. Unfortunately, the views of the contributors to the NSSE volume were not reflected in the Commission's report. This review compares the Commission's findings with the facts and observations presented in that book. Parts of this review are adapted from Fetler's (2006) discussion of that volume.

Background

The phrase, "left behind" evokes disturbing images, perhaps of an orphaned child, of a soldier wounded and trapped behind enemy lines, or of ragged survivors in a post-apocalyptic future. Actually, in the context of Public Law 107-110, to be left behind is to be on the low end of an "achievement gap." The achievement gap in question is not the one that derives from hereditary differences in cognitive abilities, between students with more or less aptitude for academics. Rather, the gap in question usually arises out of inequality in family income.

One of the best-documented and most dismal findings in educational research is the relationship between poverty and student achievement. Students in poverty typically score lower on achievement tests than their more affluent peers. (White, 1982) Moreover, poverty tends to correlate with minority and disability status. More than thirty years ago Coleman's (1966) famous report raised perplexing questions about the effects of student background characteristics on achievement. Coleman found that schools, on average, have little influence on student achievement. Non-school effects, such as social class, are overwhelmingly more powerful. The corrosive effects of poverty on learning are pervasive, affecting every aspect and hour of a person's life, and persistent, the harm done in childhood tending to cascade into later years. There will always be exceptional students, special teachers, and charismatic principals that manage for a time to defy the desolate reality of poverty. However, in the main, Coleman's observation still stands.

Unhappily, U.S. Census Bureau statistics (2005, p. 13) document persistent, growing poverty. After four years of consecutive increases, the poverty rate settled at 12.6 percent in 2005, higher than the most recent low of 11.3 percent in 2000. For children under 18 the 2005 poverty rate was 17.9 percent. While public schools are open both to rich and poor, their burden would likely be eased by economic programs that promote more prosperity for families of children in school.

NCLB's mantra of "assessment and accountability" so indissolubly bonds these two terms that it is now almost impossible to think of one without the other. However, accountability is familiar accessory to public education and there are many types, depending on who is held accountable, by whom, and for what. Fiscal accountability traditionally looks at the use of money. Federal, state, and local governments hold funded schools and districts accountable for how money is spent, 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 money for specific inputs or processes. Examples are audits, program reviews, and accreditation reviews.

Outcomes focused accountability, looks at results. How much do students learn? How many graduate from high school, get jobs, or go to college? Perhaps because it is hard to define, track and evaluate success after high school, NCLB accountability relates primarily to student achievement test scores. Teachers and schools are held responsible for raising student test scores. There is punishment for poor performance. NCLB uncritically assumes that test scores directly and faithfully reflect student learning. Like the business ethic that values greater return on investment by whatever means, if only by creative accounting, the most important criterion of educational success is higher test scores.

The Commission report traces the roots of the recent upsurge of testing and accountability in U.S. schools back to the 1983 National Commission on Excellence in Education hearings that 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 stimulate change.

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. Districts and schools must meet state-defined annual targets, make adequate yearly progress (AYP). AYP applies both to the overall student population and to various subgroups of students. 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, overall or for any subgroup, are labeled "in

need of improvement," and are subject to interventions, such as offering public school choice, supplementary educational services, and eventual reorganization.

The Commission finds signs that NCLB has not yet fully taken hold. The count of schools that have not made AYP is rising and scores on the National Assessment of Educational Progress (NAEP) are lagging. The finger of blame points at sluggish enforcement of NCLB's teacher quality requirements by the U.S. Department of Education (ED), and states' tardy implementation of testing requirements. The appropriate response, apparently, is to stay on NCLB's original course, with minor adjustments, and to extend the accountability provisions more forcefully onto teachers and principals. The Commission recommends setting up data systems and longitudinal assessment systems that track student achievement as well as teacher and principal effectiveness over time. Individuals who are deemed ineffective would at first receive staff development and eventually face restrictions on their employment.

Despite NCLB's stout support for testing since 2002, recent achievement scores provide at best tepid support for current educational policies. The National Assessment of Educational Progress (NAEP) reports long term trend data for reading and mathematics achievement for students aged 9, 13, and 17. (National Center for Education Statistics, 2007) The trend for reading achievement has been flat since 1971. Nine year olds displayed a modest gain in reading between 1999 and 2004, but it is too early to judge whether this is only a blip. The trend for mathematics achievement of 17 year olds is likewise flat. There have been modest increases for thirteen and nine year old students since 1999, but again it is premature to judge whether they are a trend.

Improved Accountability

The chapter, "Accelerating Progress and Closing Achievement Gaps Through Improved Accountability," begins, as do many of the chapters, with a story about a school. Overall, ninety percent of its students met state standards in 2004. Closer examination of results for specific student subgroups unsurprisingly revealed that less than half of special education students met standards that year. The Commission attributes the school's subsequent improvements in special education to NCLB's requirement that all such subgroups meet standards. We are to believe, were it not for NCLB, these children would be left behind. The Commission deftly concludes that the "story shows the power of actionable information and data in driving school improvement."

The report deploys stories as a persuasive technique, to illustrate an argument or point of view. In this case, the anecdote undertakes to show that accountability for specific subgroups shines a light on previously unsuspected problems and somehow produces improved achievement. Actually, the anecdote raises more questions than it answers. How many of the school's students met standards in previous years, and how many continued to meet them subsequently? What are school's policies for identifying, placing, and testing special education students? Did those policies change? What is the socioeconomic makeup of the students? Did the demographic mix of students change? Did any of the special education children leave the school, and if so, for what reasons? Was there any careful research into the causes and meaning of the changed test scores?

The report enthusiastically attributes increases in test scores to the beneficial effects of NCLB. However, a finding of this scope and importance should be supported by careful research and evaluation. This approach involves a more objective and detailed examination of the school's context and programs in relation to outcomes. For example, state and local budgets provide a given amount of funding per student. Governance of the district and school may be more or less amicable and efficient. The school is situated in a neighborhood with given demographics. Students may be transported from other neighborhoods. The physical plant may be more or less adequate to the enrollment and may or may not permit the use of modern instructional technology. The qualifications of staff and student-staff ratios have an effect on teaching and learning, as do the curriculum, student services, placement policies, and instructional programs. There are many aspects of a school's programs and environment that should be considered when trying to explain changes in test scores.

It is enticing but perilous to assume that an achievement test score perfectly reflects learning. A number with a label, e.g., "reading achievement," too easily takes on for itself the reality it claims to measure. A test score is merely a number that bears 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 psychometric properties of the test. Support for inferences about learning depend on understanding the relationship between testing and the context of schooling, staffing, curriculum and instruction, as well as anything else that could influence scores.

For accountability to be meaningful higher test scores should reflect better teaching and learning. NCLB attempts to link test scores to teaching and learning in several ways. Tests must be aligned with curriculum standards along multiple dimensions, including coverage of the full range of content in the standards; measuring 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, 2005) Perhaps more problematic to implement and verify, instruction should be similarly aligned to the standards. 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. (Choi, 2005) Unfortunately, these supports do not necessarily guarantee that scores are meaningful.

Koretz (2005) describes a threat that higher test scores could reflect coaching and not real improvement. There is a tell-tale pattern that suggests teaching to the test. Scores on a new test start out low, but show gratifyingly 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. The peaking of test scores in this situation demonstrates a version of Cannell's (1987) paradoxical "Lake Wobegon effect," where most scores are above average and do not accurately reflect student learning. 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, and test questions are representative of the curriculum, scores should 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 routinely 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 reduces opportunities 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 activities and circumstances of successful schools within each group, looking for evidence of successful programs. Although NCLB accountability relies almost exclusively on test data, deterministic formulas, and deceptively simple indexes of success, professional 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.

One likely consequence of implementing Koretz's recommendations may be slower increases in average test scores. The dramatic increases seen in "Lake Wobegon" situations can take place over a few years. Increases in scores that more truly reflect improved teaching and learning are likely to 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 the aptitude and circumstances of the student, the psychometric characteristics of the test, and the rigor of the curriculum.

The Commission's recommendation to adopt student tracking and growth models marks a significant change in NCLB's approach to testing and accountability. The Commission proposes to allow states to measure achievement growth and factor it into the AYP calculations. In this way states would receive credit for students who are making progress towards proficiency. Specifically, schools would receive credit for students who are on track to becoming proficient within three years, based on the growth trajectory of their assessment scores, when calculating AYP for the student’s school. The specific timetable comes with no justification. Three years might be sufficient for motivated and capable students who are already close to proficient, but could also be unreasonable for those who are farther away. The remedy requires that states have data systems that identify individual students, track them over years, and monitor their progress. A second requirement is that scores from a state's achievement tests be on a common scale and comparable across years.

Choi (2005) observes that the cross-sectional, status-oriented design of most state testing programs does not allow meaningful calculation of student growth. Most states have different tests for each grade, each test aligned with the content standards approved for that grade. In any given year the students enrolled in a particular grade take the appropriate test. During the course of the school year some students migrate in or out of the school, some are retained in the grade, and some promote to the next grade. The next year, once again, the students who are enrolled in a particular grade take the appropriate test. 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, the students who promote and remain enrolled in the school, and any new students, take the fourth grade test. A new group of third graders, who promoted from the second grade, take the third grade test. Even though the average third and fourth grade test scores may look similar, the tests are different, are not comparable, and do not measure growth across grades. This design only permits a report of test score trends across years within each grade level.

Interpretation of the trends is subject to the caution that the data for each year represents a different group of students, teachers, and varying context. Even though the two successive years of average third grade test scores look similar and are based on the same, or equated, test, they represent two different groups of students as well as other possible changes in teaching or administrative staff. Changes in scores may reflect changes in learning, a changing population with different numbers of poor students or English learners, changes in staff, or changing economic conditions in the community. Test systems that measure status yield imperfect estimates of growth that fail to take into account changing student or school characteristics.

Choi describes methods for estimating growth that do take into account the characteristics of students and schools. His methods involve tracking of test results for individual students from year to year and the development of a common, or longitudinal scale, that allows meaningful cross-year comparison of test scores. For example, in a cross-sectional design achievement test scores in each grade might range from a minimum of 100 to 400 points. However, a score of 350 in the third grade cannot meaningfully be compared to a score of 350 in the fourth grade. A common scale might range from a low of 100 in the third grade to a maximum of 500 in the fourth grade, and scores could be compared across years to estimate growth in achievement. If the tracking includes information about teachers and schools, it is possible to calculate not only individual gains, but also teacher and school growth. Testing designs that allow these calculations of growth along with adjustments for student or school demographics are called value-added models.

A common scale depends on the existence of a logical continuum across grades of the tested curriculum. One dimension of this continuum may reflect the increasing complexity and breadth of academic language required as students grow older and progress from grade to grade. Another dimension may relate to the scope and logical sequence of material that is taught. Because tests must be aligned with content and achievement standards, those standards should also reflect a logical continuum across grades. It is fair to ask whether states sufficiently consider these kinds of requirements for continuity when they write their content

standards and set their achievement standards. Of course, school curriculum usually reflects some consideration of the scope and sequence of material across grades. However, the question is whether that consideration suffices to support the technical requirements of a common scale.

The Commission recommends changes in the way that results for subgroups are factored into accountability calculations. NCLB now requires that each year every subgroup must increase the percent of students who are proficient, culminating in 100 percent proficiency in 2013. A school falls into improvement status if during a two year span any subgroup does not make AYP. That is, during the first year one subgroup might fail in reading and during the second year a different subgroup might fail in mathematics. If so, the school is identified for school improvement. The "any-subgroup" requirement quickly ratchets up the proficiency requirements, making it difficult for schools to make AYP. The recommendation would only identify schools for improvement if they do not make AYP for the same subgroup in the same subject for two consecutive years.

It is difficult to say whether the "same-subgroup" modification makes much difference in the difficulty of making AYP. The reasons are that students can belong to more than one subgroup, and membership in some subgroups is correlated with membership in others. For example, students in poverty are more likely to be minority, English learners, or have disabilities. To the extent that membership in subgroups overlaps, failure must also overlap.

Student tracking would enable other refinements in the application of AYP to subgroups. The "same-subgroup" approach could be sharpened by only including students who remain in a given cohort. Students who have been enrolled for two years would be accountable for making two years of growth. Those enrolled for three years would be accountable for three years of growth, and so on. Another modification would be to count a specific student in only one subgroup. Consider the Hispanic student who is an English learner, is in poverty, and has a learning disability. In the current system this student potentially counts four times for or against the school, depending on increases or decreases in his or her test scores. Perhaps this student should count in each of the four subgroups. The advocates for English learners wish to claim him or her, as do the advocates for disabled students. However, an alternative is to count the student in just one subgroup on the principle of one person, one vote.

Student Progress

"Fair and Accurate Assessments of Student Progress" is the title of the chapter in the Commission's report that reviews how states have implemented NCLB's testing requirements. Citing the results of the U.S. Department of Education's review of state testing programs as of July 2006, the report notes certain major concerns, namely, that states need to: demonstrate that alternate assessments for students with severe disabilities are comparable to regular assessments; provide appropriate accommodations for students with disabilities and English language learners; demonstrate the alignment between tests and standards; and demonstrate that the results of different forms of assessments (paper-and-pencil tests, computer-based assessments, assessments translated into Spanish) were comparable. NCLB also requires reports of performance on tests that allow parents and teachers to understand and address the

academic needs of students. Despite the requirement for clear and timely information, there are complaints that the reports are filled with arcane technical language. In general, the Commission recommends continuing federal funding to improve the quality of state assessments.

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. Testing accommodations permit the inclusion of those students in accountability programs that might otherwise not participate. The idea of fairness in education at one time was linked primarily to the belief that all students should have an equal opportunity to learn. NCLB stretches this blanket of fairness to include the thought that standards, testing, and accountability should also apply equally to all students. In the past students with disabilities or English learners did not necessarily participate in regular testing and accountability programs. The goals and teaching for these students were different, leading to their exclusion, 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 now 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. However, changes in the standard conditions of testing cast doubt on the validity of test scores. 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? How much does reading a test of reading comprehension test to a blind student, or signing to a deaf student, change the meaning of the scores?

Pullin (2005) states that the question for students with disabilities is not whether, but rather how, they take part in testing. 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. Ideally each state sets forth general procedures for accommodations, and the local IEP team makes informed decisions, depending on the specific needs of the individual student and the characteristics 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 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 often lack the information and expertise needed to make appropriate decisions.

Pullin notes the meager research on accommodations and alternate assessments. 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. English learners, as well, have varying degrees of English language proficiency, different language and cultural backgrounds, and differing amounts of literacy in their primary language.

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. The Commission recommended developing alternate assessments for English language learners. While an alternate assessment for some English learners may improve meaningful participation in state assessments, Abedi (2005) asserts that correct identification of English learners 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.

Tests that are given in English measure not only performance, 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 threat to validity becomes more serious as the language demands of the test increase and as levels of English proficiency go down. Accordingly, using simplified English in the directions for administering tests, and in test questions when the topic allows, may be an appropriate accommodation. Another 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 cultural context. If the student receives instruction mostly in one language, a better option may be to give the test in that language. Other possible accommodations are access to a glossary, or extra time.

An alternative to accommodations is to provide "universally designed" tests that have fewer barriers to participation, for example, by 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 efficiency, the reliability of test scores, and eases interpretation of results, 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 inevitable decreases in the reliability and validity of test scores.

Effective Teachers

The chapter titled "Effective Teachers for All Students, Effective Principals for all Communities" discusses ways to extend NCLB accountability for student achievement to school staff. Historically, more experienced and qualified teachers teach at schools serving more affluent students, while urban schools that serve low-income students tend to be left with newer, less qualified teachers. NCLB's "highly qualified teacher" (HQT) requirements are intended to redress this imbalance. NCLB requires that "highly qualified teachers" possess state certification or licensure, have at least a B.A. degree, and demonstrate knowledge of the subjects they teach. The Commission recommends tightening teacher requirements by additionally demanding proof of "effectiveness." States would be required to establish systems to measure the learning gains of a teacher's students through a value-added methodology, using three years of student achievement data, as well as principal evaluations or teacher peer reviews. Teachers who fall in the top 75 percent of statewide learning gains and receive positive evaluations would receive "Highly Qualified Effective Teacher" (HQET) status. Teachers who do not receive HQET status would be subjected to additional professional development, parent notification, and eventually could not be employed in Title I schools.

Heritage and Yeagley (2005) and Shaw (2005) look at classroom practice. They note that 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. Linda Darling-Hammond and Elle Rustique-Forrester (2005) underscore the point. Negative consequences can result from accountability systems that do not attend to improving teaching. 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.

In theory, value-added models permit an evaluation of a teacher's contribution to student achievement. However, if such methods influence personnel decisions and limit employment opportunities, they may also provoke vigorous dissent. For example, schools vary in the amount of funding provided per student, in the quality of their facilities, and in the types of curriculum they provide. How much do these factors influence achievement, and how well are they accounted for in value-added models? How much influence does a teacher have, compared to the influence exerted by student demographics, culture, language, disabilities, and poverty? If a teacher's influence is relatively small compared to these other factors, does it make sense to use a "top 75 percent" criterion, as proposed by the Commission?

It is rash to assume that a precisely stated goal based on test scores is in fact precisely measureable. The statistic "top 75 percent of statewide learning gains" looks like a simple and pure measure of performance. However, test statistics, like the truth, are rarely simple and never pure. Every such statistic contains a measure of error that stems from the conditions of testing. That error certainly produces misclassification of a larger or smaller number of teachers, depending on the statistic. Rogosa (2005) writes about the statistical properties of the scores that NCLB accountability programs use. He discusses 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 example, calculating the percent of students who are making learning gains. Judging involves making a decision, depending on a measure, whether a teacher met criteria for success. The challenge is to estimate the amount of error in those measures, and to make judgments in a way that is statistically sound and professionally acceptable. While there are statistical rules of thumb for simple situations, measures for school accountability are complex, and call for the professional judgment of statisticians. Rogosa's caution certainly applies to statistics used in making personnel decisions, as well as those used to make decisions about the improvement status of schools.

The Commission report noted NCLB's requirement that reports of test results allow parents and teachers to understand and address the academic needs of students. The annual achievement tests administered at the end of the year in order to evaluate AYP serve the interests of neither parents nor teachers very well. Teachers need detailed information about student strengths and weaknesses during the course of the school year in order make adjustments to instruction. Parents need reports on their children's performance during the year, and students should know how they are performing so that they can focus their efforts appropriately. It is not clear that the statewide tests administered for NCLB school accountability can adequately or equally well serve these different interests.

Different tests are suited to different purposes. Benchmark tests are given by school districts several times a year to monitor the effectiveness of instructional programs. These tests might cover several months of instruction in detail. Teachers administer classroom tests to monitor individual learning. These tests typically cover shorter intervals of time and reflect the curriculum taught by a particular teacher in a particular class. To the extent that they are psychometrically sound, classroom tests can yield snapshots of each child's progress in a course. The frequency, customization, and detail of classroom tests better suit them to instruction and diagnosis of individual needs than to the evaluation of schools or programs. 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. Summative tests tend to yield general information about groups of students useful for program evaluation. The high quality tests created for the National Assessment of Educational Progress (NAEP) provide a good example of a good summative assessment that does not provide useful student reports. NAEP only tests a small random sample of students. Each student only responds to a small sample of the items included in the entire test. Although NAEP does not produce individual student or school results, it efficiently produces high quality reports of state and national performance.

Baker (2005) suggests ways in which testing can better serve teaching and learning. Her approach distinguishes between structural and functional approaches to accountability. Structural methods for improving school focus on distinct parts of a system (standards, teaching, tests, flow of information, incentives) and their alignment. By contrast, functional methods focus on the people in schools (students, teachers, managers), their roles (learning, teaching, coordination), and how they interact to accomplish their goals. Baker's more functional 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.

Baker 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 underpin 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. 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?

Comparing States

The chapter entitled "High Standards for Every Student in Every State," discusses the large differences in content and achievement standards across states. Each state implements a different and unique set of achievement tests designed to assess its own specific content standards. Each state's tests are scored using its own achievement standards that are aligned to its content standards. The variety of content and rigor mean that no state's standards can be directly compared with any other states standards. The achievement test scores of any one state cannot be directly compared with the scores of any other state. The report outlines three

recommendations. First, states should assess their reading, mathematics, and language arts standards against requirements for success in college and demanding jobs, including existing national and private efforts to identify college and workplace readiness skills. Next, a national panel should develop a voluntary model, based on NAEP frameworks, national content and performance standards and tests in reading, language arts, and science. While the NAEP frameworks are to be used as starting points, the panel should see to it that the voluntary model would sufficiently prepare students for higher education and jobs. Finally, the U.S. Secretary of Education should periodically issue reports comparing the rigor of all state standards to the national model standards and tests using a common metric.

Linn (2005) also notes the variability in state's accountability systems. Historically, states assessed different subjects in different grades and attached different rewards and sanctions to the results. There has been a general trend to raise the stakes in accountability systems that in the past may have only encouraged students and teachers to work harder. 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 performance standards was to make it easier for the public to understand reports of test results and to set expectations for acceptable levels of proficiency. Unfortunately, the great variety in performance standards across states has made it more difficult to understand those reports.

The National Assessment Governing Board (2002) (NAGB) reported on the potential use of NAEP to confirm state test results. The Board found that NAEP could be used as evidence to confirm a general trend of state results in grades four and eight reading and mathematics. However, limitations in using NAEP should be acknowledged, meaning that the confirmation is not purely statistical and depends on a "reasonable person" standard. These limitations 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 board's report appears to support the use of NAEP as a single point of comparison for states. However, the comparison must be qualified by the numerous differences in testing practices across states and NAEP. The limitations in making comparisons to NAEP would likely further complicate attempts to make direct comparisons of one state with another.

Stake (2007) writes that the original purposes of NAEP were to further educational research and to authentic education reforms resulting from the war on poverty and the cold war. Its intended original role is far different from its recommended part now as the overseer for state level implementation of NCLB accountability. If the logic of NCLB is correct that tight alignment of testing, teaching, and standards produces desired learning, does it not suggest that a more authoritative voice for NAEP encourages alignment of testing and teaching to its own national content standards?

The Commission's interest in college and workplace readiness focuses on whether group administered achievement tests can reasonably be used to predict future success of individual students. By design group achievement tests are intended to portray in broad strokes the results of past teaching and learning. Group administered tests, developed for school and

district accountability, may provide reasonably reliable and valid aggregate scores, but rarely report detailed, reliable, diagnostic information about individuals. Such tests are usually evaluated primarily for their alignment with the state's specific content standards and the reliability of their scores.

Predicting the future is different in that success at work or in college is based on the mastery of tasks that differ from those taught and learned in school and depends in part on more general skills and abilities than those tapped by achievement tests. The job of predicting future success traditionally has been the domain of aptitude tests, such as the GED, the SAT or the ACT. These tests tend to be more general in their content, take longer to administer, are developed to assess broader cognitive skills, provide some diagnostic information, and are evaluated with regard to their success in predicting future performance. If achievement tests administered for NCLB accountability are to take on the additional task of predicting individual future success, these tests must take on more of the characteristics of the GED, the SAT or the ACT, including increased administration time and expense. Few state assessment programs are now capable of taking on this task.

Retrospection

NCLB does little to encourage schools and states to reflect on the reasons for changes in test scores, and there is little in the work of the Commission to suggest a change for the better. Educators often trumpet any gain in test scores, however miniscule, as evidence of significant improvement. There is political value in reporting good news, and such reports may help to rally the faithful. However, genuine progress in teaching and learning is likely better served by a genuine understanding of the situation.

There are enough stories about mechanical failures in statewide testing programs to warrant improved quality control. Schools and districts need evaluations to discourage improper test preparation and administration. Procedures for test scoring and reporting are complex and prone to error. Student tracking and value-added models multiply this complexity by orders of magnitude. 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 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. Quality control audits provide necessary checks on the work done by test contractors.

After establishing the integrity of the results, further 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 steady upward trend may be trivial if it takes decades to reach a goal.

One reason for the relative lack of research and evaluation of test results may be the scarcity of appropriately trained staff. Even large external testing programs run by large companies often operate with only a few higher salaried technical experts who design, implement, and monitor tests, sometimes for many different states. While technology enables large-scale printing, distribution, scoring, and reporting of tests, the staffing is stretched thin for appropriate quality control and evaluation of the results.

NCLB embeds testing deeply into teaching and learning, but neglects to examine the consequences of this intrusion. 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.

NCLB massively funds testing and accountability, but most of money goes into printing 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. The U.S. Department of Education now funds a NAEP staff liaison at each state department of education. These staffers are dedicated to NAEP and receive extensive training in relevant assessment topics. One way for NCLB to strengthen state offices would be to implement a similar program. 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.

The Commission report advocates systems for tracking students and value-added testing models. While these changes technically improve the status-oriented systems implemented by most states, they are not a panacea. High student mobility may undermine this technology. Some states are experiencing high rates of immigration by both documented and undocumented individuals. For instance, gateway schools near the border between the U.S. and Mexico have a large turnover in the student population each year. Migrant workers make up larger percentages of student enrollment in certain agricultural areas and contribute to school mobility. It is also possible that students in poverty or single-parent homes are more mobile for reasons associated with their socioeconomic status. The identification and tracking of these students may raise legal questions relevant to citizenship status, child support, taxes, and privacy. Implementation of student tracking data systems becomes more difficult to the extent that these ancillary issues appear inconsistent with the goals of education or in some way threatening to the well-being or rights of families. Perhaps research with student tracking and value-added models could clarify the relative effects of poverty and schools on student achievement and what amount of genuine improvement is realistic.

The implementation of data tracking systems makes possible a more detailed analysis of student mobility. The movement of students from school to school, district to district, or

across state lines each year decreases the size of the cohort that remains in a given school or district over time. Schools and districts with higher rates of mobility may find that few of the originally enrolled students remain after several years. Holding both low and high mobility schools accountable to the same schedule for AYP raises questions of equity.

The report does not explore potential limitations of common scales to measure growth across more than a few grades. Teaching goals, methods, and materials change across grades to reflect student growth and prior learning. The tests must reflect these changes. After several years it becomes less meaningful to compare scores across grades. For example, it makes little sense to compare first grade reading performance that emphasizes phonemics with sixth grade reading that emphasizes paragraph comprehension. High school algebra and geometry do not easily compare to arithmetic taught in elementary grades.

The Commission voices concern that states vary widely in their approaches to NCLB testing and accountability. State content standards vary in their level of detail and rigor. Tests vary in their content and difficulty. Schools in some states may find it easier to make AYP than schools in other states. The diversity is unavoidable because states with their diverse history and priorities, not the federal government, have primary responsibility for public education. States, not the federal government, provide the bulk of funding for schools. State education systems vary across many important dimensions, including the amount of funding annually provided for each enrolled student, funding for facilities, student-teacher ratios, the required amount of teaching, textbooks, and so on. A systemic view of education would consider that all of these dimensions are related. Attempting to make testing and accountability more similar across states might have the effect of encouraging more similarity in other ways. Leveling of systems would be more likely if testing and accountability actually function to drive other components. However, if funding, staffing, textbooks and facilities are more powerful drivers, the attempt to homogenize testing and accountability might only be frustrating and have little effect.

NCLB's menu of tests includes NAEP, which inevitably draws the eye, yet with ambivalence, just as a rich dessert tempts us after a full meal. We expect it to taste good, but fret that it is not good for us. The NAGB report cedes a role for NAEP with one hand, but takes away with the other, citing numerous cautions and qualifications. The Commission seems to respect NAEP's technology, but questions its battlefield rigor and suitability for assessing the advanced skills needed for jobs and college. Stake praises NAEP for its psychometric prowess, but observes that it is miscast as an accountability overseer. As political interest in assessment and accountability continues, NAEP may well continue its double role, both as an armament in NCLB's political war, and as a scientific instrument for measuring educational outcomes.

Ultimately, how will the Commission influence the reauthorization of NCLB? Lorraine McDonnell (2005) observes that NCLB is a political response to perceptions of the achievement gap. Politicians believe that testing provides information and leverage for holding schools accountable and motivating them to be more responsive to parents and taxpayers. High-stakes testing links the results to penalties 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 material benefit. Business groups support testing as a way to increase the efficiency of hiring good workers, and 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. Public education appears to have a public relations problem.

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 teaching materials. Considering the perceived advantages, politicians may very well stay the course on standards and testing.

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.

Baker, E. (2005). Technology and effective assessment 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.

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.

Choi, K., Goldschmidt, P., Yamashiro, K., (2005). Exploring models of school performance: From theory to practice. 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.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of educational opportunity, Washington, DC: U.S. Government Printing Office.

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

Darling-Hammond, L., and Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. 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.

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

Heritage, M., and Yeagley, R., (2005). Data use and school improvement: Challenges and prospects. 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.

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.

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.

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 April 30, 2007 from http://www.nagb.org/pubs/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.

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.

Rentner, D.S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S. and Zabala, D. (2006). From the Capital to the Classroom: Year 4 of the No Child Left Behind Act. Washington, DC: Center on Education Policy.

Rogosa, D. (2005). Statistical misunderstandings of the properties of school scores and school accountability. 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.

Shaw, J. (2005). Getting things right at the classroom level. 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.

Stake, R. E. (2007). NAEP, Report Cards and Education: A Review Essay of Jones, L. and Olkin, I. (2004). *The Nations Report Card: Evolution and Perspectives*. *Education Review, 10*(1). Retrieved July 7, 2007 from http://edrev.asu.edu/essaysv10n1index.html.

White, K. (1982). The relation between socioeconomic status and achievement. *Psychological Bulletin, 91*, 461 - 481.

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