



The Gordon Commission
on the Future of Assessment in Education

Testing Policy in the United States: A Historical Perspective

Carl Kaestle
Brown University

The author acknowledges very helpful expert suggestions, based on an earlier draft of this essay, from Michael Feuer, Frederika Kaestle, William Reese, Lorrie Shepard, and Tracy Steffes. Also, thanks to Elizabeth Hollander and Alexa LeBoeuf for stylistic suggestions, as well as research assistance from Ms. LeBoeuf. The remaining substantive and stylistic flaws are all mine.

The audience for this essay

This essay presents a history of educational testing in the United States, with an emphasis on policy issues. A number of excellent pieces have already been written on this subject.¹ It is therefore appropriate to address two questions: Who is the intended audience for this essay, and what is its intended purpose? The invited papers of the Gordon Commission on the Future of Educational Assessment will be published and will, hopefully, reach a larger audience that includes other education researchers, teachers, school administrators, and policy makers from the local to the federal level. Therefore I do not expect my readers to have very much prior knowledge about this subject, and I apologize to those expert readers who already are familiar with the details of this history. The intended purpose of the essay is to reflect upon how present testing practices developed and to spur ideas about how we might use assessments more in keeping with the educational values and needs of the twenty-first century.

Testing in nineteenth-century America

Histories of testing often start by describing written civil service examinations in China, which were implemented around 200 B.C.E. From this notably ancient starting point, the story leaps to the introduction of written exams at the English universities in the eighteenth century. These then spread widely across Europe.² As important as these precedents may be, I shall begin with the early history of testing in the United States, a story that is not so well known. While some writers make a slight nod to nineteenth-century written exams, the focus quickly fastens on the introduction of intelligence tests in the early twentieth century, moving on to the widespread adoption of standardized, multiple-choice tests in the 1920s and beyond.

That view implies that the history of testing encompasses only tests that were uniform, one-shot exercises, comparable across classrooms and across schools, used for high-stakes decisions like grouping, promotion, admission, or evaluating schools' performance. But that emphasis leaves out two sorts of assessment that long preceded standardized tests. First, teachers made judgments every day in classrooms, asking children to practice reading, arithmetic or other subjects out loud, "correcting" them, telling them how to improve their performance, and estimating how well individual

students were mastering the curriculum. It was a very oral environment; not much writing occurred. Teachers lectured, students recited. Documentation is sparse for these informal assessments devised by classroom teachers; they are difficult to research historically, but we should bear in mind, as Michael Feuer has commented, that classrooms are complex settings that “require nearly constant ‘scanning’ by teachers,” the best of whom display “a willingness and ability to integrate inchoate information rapidly.”³

Many reformers today urge that we redress the balance between standardized tests with national norms and local assessments that are integrated into teaching and learning. Indeed, in the late eighteenth and early nineteenth centuries, before standardized tests came on the scene, teachers did most of the assessing through mundane classroom procedures. That does not make the good old days of rote learning and recitation a model for the future, but the comparison reminds us that the dominance of standardized tests was not inevitable. It was historically conditioned.⁴

Time has also obscured a second type of assessment. A century ago teachers prepared their students to perform periodic exhibitions for the school committee or, more often, for the community as a whole. Teachers spent weeks preparing students to show off their skills at oratory, singing, memorization and other aspects of the curriculum. Like daily recitations, exhibitions served several functions. For parents it was an entertaining occasion, a chance to cheer proudly for their children, and a chance to assess how accomplished the teacher was. Although students surely had anxieties about their performance, exhibitions were not designed as assessments of individual children. Schoolmasters decided who would be promoted to higher levels of schooling on the basis of daily classroom performance.

Beyond the traditional recitations and exhibitions, “standardized” tests became more popular in the second half of the nineteenth century. These were tests initiated by authorities beyond the classroom teacher and administered uniformly across a school or a district. Fortunately, we now have a comprehensive work on school exhibitions and the advent of written examinations in nineteenth-century schools in the United States, by historian William Reese. His main points, briefly suggested here, are beautifully documented in his forthcoming book, *The Testing Wars*. They are relevant to understanding our current testing policies and their relationship to teaching and learning.⁵

Horace Mann was Secretary to the Massachusetts School Board. In 1845 he and his reformist allies reacted against the system of exhibitions and the lack of accountability of schoolmasters. The reformers championed periodic written exams, to be given at least annually, across all the grammar schools of Boston. Mann had visited Europe extensively and published some admiring observations in his Seventh Annual Report. The testing proposal was but one part of a reform agenda that included hiring more female teachers, training them in a more humane pedagogy at normal schools (teacher training institutions), installing a superintendent to oversee the city's schools, and using written test results for promotion and for the re-contracting of teachers. Most of the enthusiasts were members of the Whig Party, which favored centralized government involvement in public works and social institutions. They were also devotees of the new science of statistics as a tool in such endeavors.⁶

The Boston schoolmasters and their supporters on the School Committee opposed the move as a threat to their autonomy, and, indeed, the reformers saw their testing plans largely as a way to assess teaching quality and to use the externally-produced testing data in making decisions about the re-contracting of teachers. During a tumultuous battle that stretched on for several years, the Mann phalanx gradually prevailed. Exams were established, a superintendent hired, and a normal school established. The testing advocates on the Boston School Committee published a lengthy report on the testing program in Boston, along with analysis of the first test results. The report was widely praised in the fledgling education journals of the day. These practices spread from city to city, from the Northeast to the Midwest and the South. Tests proved popular in cities because large school systems raised questions of comparability and standards. They also needed more systematic methods of classification and placement than small rural schools. In addition, as public high schools were established in cities, they devised admission tests, a beachhead for educational testing.⁷

Exhibitions, recitations, and spelling bees persisted, especially in rural areas, but after the Civil War, Reese writes, those cultural survivals existed in “a new educational landscape.” Immigration, urbanization, railroads, and national markets for agriculture enhanced the use and popularity of statistics for depicting and analyzing social phenomena. William Wells, superintendent of Chicago's schools from 1856 to 1864, saw

an analogy between educational tests and thermometers and gauges in industry. They provided a way for him to centralize and enhance his authority over the system. As in 1845, there were critics of tests, and their voices grew louder as the tests became more prevalent and time-consuming. As more school districts established graded levels of classes, textbook publishers provided grade-level texts. Soon tests were available for grade-to-grade promotion and then manuals and test-prep books to match. Teachers still made the decisions, but tests were becoming part of the system in many school districts.⁸

School superintendents, including those in Omaha and Portland, applauded the usefulness of tests in sorting out children entering their systems from un-graded schools, either public or private, or migrating from other towns or from abroad. Superintendent A. J. Rickoff in Cincinnati developed elaborate rules for test administration, the prevention of cheating, and the modes of scoring. The spread of written tests took place in the context of a gradual but salient trend in schoolroom pedagogy: the withering of oral traditions and the increase in writing. No more was writing confined to copybooks and slates. Now there were notebooks, essays, improved pens and pencils, blackboards, and written assignments.⁹ Thus, one could see the advent of more written tests as an adjustment in tandem with changing technologies of pedagogy. Nonetheless, acknowledged or not, the shift also raised political, ethical, and philosophical questions.

As tests became more pervasive, administrators and teachers felt the weight of endless hand-scoring of tests. Critics complained that tests encouraged a heartless, stifling pedagogy. *Harper's Magazine* and the *Atlantic* published articles decrying the triumph of memorization over understanding. Some claimed that tests were ruining students' health and causing suicides. But it was hard for administrators to imagine teaching methods that would respond to each child as an individual in urban systems with burgeoning immigrant populations, scantily prepared teachers, and classrooms with 50 to 60 children in the primary school years. A majority of urban school board members favored the reform regime, including testing. As for rote learning, it was an old tradition, and now, writes Reese, memorization was "not only alive and well but aligned beautifully with the new culture of competitive testing." Memorization skills that had been good for traditional recitation proved essential to silent test-taking as well.¹⁰

As Reese points out, the solution to a criticism of a particular type of test was often to expand the use of other tests. Thus when parents and other critics focused their complaints on the unfairness of one-shot grade-promotion tests, administrators phased them out in the 1890s and promoted frequent testing on a daily, weekly, and semester basis. And when the progressive educator Francis Parker complained about the mania of tests, one of his critics pointed out that the students in Parker's normal school were two years below grade level. By this time, it seems, the criteria were installed, the language established. In 1900 William Powell, the superintendent of schools in Washington, D. C., testifying before a Congressional committee, was berated for opposing tests and thus lowering standards. He was fired.¹¹

And thus the century ended. Reese draws a stark conclusion from this tale: "Anyone who thinks that recurrent attacks on high-stakes exams will lead to a diminution in the number and authority of tests is surely mistaken." By 1900 the research and experimentation on intelligence testing had begun, adding another accelerating breeze to the enthusiasm for testing.

Intelligence testing: The early years

The mid-nineteenth-century rise of educational testing occurred on the cusp of an era of influential racial and biological theories about behavior and capacity. It is not surprising, then, that many innovators in education testing became interested in the concept of general intelligence. The first generation of professional psychologists who tackled that problem of measuring intelligence were a varied group, with different backgrounds, aims and theories. Sir Francis Galton, an Englishman and one of the elder savants, was raised as a child genius by parents who were confident that the boy's keen intelligence was inherited. Galton became a lifelong hereditarian. He established an "anthropometric" laboratory in London. Most anthropometric laboratories up to that time emphasized differences in height and other physical traits, but Galton emphasized differences in mental abilities between individuals and groups of individuals. This soon led him to ponder ways of dealing with correlations, spurring a line of early psychometrics research that was pursued by two of his protégés, Karl Pearson and Charles Spearman.¹²

It is worth highlighting here that during this period, just as in Horace Mann's day and in our own time, American researchers were engaged in trans-Atlantic networks of innovation in testing and education reform. As cosmopolitan as that may seem, something about the American context led the United States ultimately to place much more emphasis on standardized tests than European educators. Perhaps it was the great size of the country and its federalist structure of fifty states, each having strong prerogatives regarding education governance. Given this tradition, some educators and school board members felt a need for national comparisons and measures of accomplishment amidst a chaotic, loosely-coupled system featuring local control by lay boards, in contrast to European and Asian "ministry" systems, which often had common curricula and examinations at a national level. A comparative history is beyond the scope of this essay, but, as we shall see, America's different development is a policy question of considerable importance.¹³

Galton's American protégé, James McKeen Cattell, continued Galton's anthropometric approach to mental traits at Johns Hopkins University, the University of Pennsylvania, and Columbia University. However, the physiological approach to differences came under scrutiny, and critics impugned the correlational conclusions of Galton and Cattell. Nonetheless, the genetic emphasis in Darwinian theory held sway, and the convictions of the pioneer generation lived on. Many psychologists of the early twentieth century continued the quest to define "g," a person's general or summary level of intelligence. They believed that intelligence thus construed was hereditary, unitary, and largely immutable.¹⁴

Alfred Binet, however, took a different approach. Born in Nice, France of well-to-do parents, Binet had a depressing childhood, dominated by a stern father, a physician who once tried to force Alfred to touch a cadaver in order to toughen him up. Wandering from profession to profession and enduring an emotional breakdown, Binet began reading on his own at the Bibliotheque Nationale in Paris. He found his calling in psychology, with a focus on the individual's life course and emotions. After some false starts, Binet developed a methodology of case studies and an interest in developmental stages. Both reinforced his individualistic and environmental convictions about intelligence.

A series of coincidences pushed Binet's career into work that attracted worldwide attention. In 1899 Theodore Simon, a young physician interning at an institution for mentally subnormal children, applied to work with Binet for a doctorate, giving the forty-three-year-old Binet access to this new and interesting population. This coincided with the French government's decision to require schooling for all children, including special classes for those with mental limitations, and in 1904 they selected Binet to identify and characterize childhood learning deficiencies.¹⁵

Pursuing his long-standing conviction that age was a central factor in intelligence, Binet discovered that mentally subnormal children could often accomplish many tasks that normal range children could do, but only at a more advanced age. This led to a strategy of testing children with a range of tasks of increasing difficulty, with various thresholds becoming cut-points for classifying children as idiots, imbeciles, or morons. The scales he produced in 1908 and 1911 (the year of his death) were thus much more pragmatically oriented than others' descriptions, even though his concept of intelligence was more multiple. Binet thought that intelligence was a constellation of different skills, not a single trait.¹⁶

Binet died young in the midst of this exciting work. He had not developed a wide following, and he had not trained many protégés. Those who were impressed by his approach and the practicality of his scales tended to interpret them in hereditarian terms, despite the fact that Binet had not thought intelligence was innate or immutable. Henry Goddard, who had studied with G. Stanley Hall at Clark University, translated the Binet scale into English and tried it out on a diverse group of retarded children at his Training School for the Feebleminded at Vineland, New Jersey. It seemed to work well at predicting the classifications of the children, and by 1916 Goddard had become one of the leading proponents of IQ testing in the United States. He was steeped in the hereditarian tradition, even more absolute than Galton. His 1912 book, *The Kallikak Family: A Study in the Heredity of Feeble-Mindedness*, promoted strong eugenicist policies, such as institutionalization and sterilization, to prevent a scourge of feeble-mindedness from enveloping the United States. *The Kallikak Family* argued, on the basis of ambiguous data, that feeble-mindedness was genetically caused, that it controlled behavior and capacities, and that it was nearly impervious to change.¹⁷

Another of Hall's students boosted the Binet intelligence concept into a widespread testing business in the United States. Lewis Terman went to normal school to become a teacher in Indiana, then a principal. He eventually got a doctorate at Clark University, studying under Hall. There he became passionately involved in issues of educational measurement and general intelligence. Imbibing none of the skepticism and caution of his mentor, Termin wrote his thesis on "Genius and Stupidity," studying seven "bright" boys and seven "stupid" boys. He became a professor at Stanford, under the deanship of Ellwood Cubberley, one of the key cheerleaders of the educators who promoted efficiency, classification, and testing, those whom David Tyack has called the "administrative progressives."¹⁸

Termin's biggest achievement was a revision of Binet's test of general intelligence. He adapted and then popularized the term "intelligence quotient," a score obtained by dividing the subject's mental age by her chronological age. By 1916 Termin had produced the Stanford-Binet intelligence test, recalibrated from Binet's scales to better match the U. S. population. He was now the leading innovator in the IQ enterprise, which brought him not only fame (president of the American Psychological Association) but fortune. He was also a eugenics enthusiast, favoring immigration restriction and sterilization of low IQ people to save society from the "menace of the feeble-minded." Although Termin was less cautious than some fellow psychologists, he was preaching to a large, receptive choir of people who believed in the basic tenets of the eugenics movement, that IQ was hereditary, that it was lower among the Eastern European and Mediterranean immigrants, and even lower among black people. For people with these beliefs, limiting the reproduction of low-IQ individuals seemed a reasonable public policy.¹⁹

IQ testing in the era of World War I

The publicity surrounding the widespread use of intelligence tests to classify army recruits during World War I boosted the prestige of these new tests. However, behind the scenes, many Army officers were skeptical and resisted the testing program. Within the high command, officers resented the incursions of several hundred civilian psychologists, usurping military decisions better made by military officers who knew the nature of the

jobs to which the men were being assigned. They complained that assigning a permanent intelligence score to recruits with limited English ability was neither justified nor productive. The officers argued that non-English-speaking recruits should be given some English lessons, whereupon some would become able and accomplished soldiers. At the end of the war, the Surgeon General and the top brass shut down the testing operation over the objections of its civilian director, Robert Yerkes.²⁰

Yet for several reasons the public school establishment welcomed the new mental tests. As we have seen, urban schools were already using tests to assist in placement decisions, for which many educators found intelligence tests promising; and these practices were spreading out into the smaller towns. Second, the values of efficiency and scientific decision-making held even more powerful sway in early twentieth-century America than in the nineteenth century. Horace Mann hewed to the science of his day, in his case phrenology and the fledgling discipline of statistics, but early twentieth-century education theorists brought to a new level the idea of science in the service of efficiency.

Even before the publicity surrounding the military testing of recruits, intelligence testing was an exciting prospect to educators. J.C. Bell wrote in 1912 that “no device pertaining to education has ever risen to such sudden prominence in public interests throughout the world.”²¹ The popularity of new tests was a response to demographic and economic shifts that dwarfed those of the mid-nineteenth century. Immigration rates tripled; those entering from Northern and Western Europe in 1882 had represented 87 percent of all European immigrants, in contrast to 13 percent from Southern and Eastern Europe. By 1907, those percentages had nearly reversed: 19 percent from Northern and Western Europe and 81 percent from Southern and Eastern Europe. These immigrants were poorer and more “foreign,” that is, diverging even more in religion and ethnic histories from the bulk of the existing population than nineteenth-century immigrants. The scale of manufacturing enterprises ballooned, national brands and corporations multiplied, while rail networks and telecommunications knit the country into one bustling marketplace.²²

Historians have often remarked on the “search for order” in this period, the “incorporation of America,” the “managerial revolution,” the “control revolution,” and “efficiency and uplift” in the scientific management movement. Government expanded

along with corporate enterprises, and the management of large bureaucratic institutions became an unprecedented priority. Handbooks, manuals, rules, schedules, plans, and “time/motion” studies became the stock-in-trade of a whole new class of white-collar workers: managers, planners, contractors, middle-men, accountants, financial workers, lawyers, architects, engineers, and clerks.²³

School managers were drinking from the same trough. Reese’s work depicts the early formation of modern urban school systems in the 1880s and 90s. The chief organizers and boosters were urban school superintendents, occupying a new job title invented in the late nineteenth century. They self-consciously modeled their school systems on the business corporation, using tests to monitor their systems and make placement decisions for ability grouping in the lower schools and for tracking students into the newly differentiated curricula of rapidly expanding high schools. As the percentage of students age 14—17 going to high school expanded, the schools began offering vocational and commercial courses for the non-college-bound. The school became the arbiter between the family and the economy.²⁴

Paper-and-pencil intelligence tests that could be administered to groups proliferated in the late 1910’s. Prominent among these was an effort by the National Research Council to produce a National Intelligence Test. In 1919 they convened five prominent psychologists (including Yerkes, Cattell, and Thorndike) to devise a test especially for school use. Dozens of alternative forms of the examination were produced in order to make coaching less likely. Thorndike predicted that schools in all large cities would be developing classes for gifted children and that the parents of those children would be striving to coach them for their intelligence tests. In 1920, the first year this “National Intelligence Test” was available, it sold 200,000 copies.²⁵ After the war the multiple-choice IQ test developed by Arthur Otis, a member of Yerkes’ team, was marketed by *World Book*, the encyclopedia publisher.²⁶

Critics of intelligence testing

The certitude of most early IQ enthusiasts about the heritability and immutability of intelligence was not warranted by the facts and was impugned by critics at the time. Several psychologists were skeptical from the beginning, and some did extensive

demonstrations that the IQ enthusiasts' research was factually flawed. The most celebrated critic, however, was Walter Lippmann, the liberal columnist, who wrote a series of attacks on the claims of Terman and other IQ enthusiasts for *The New Republic* in 1922. Lippmann argued that the test questions did not necessarily represent "intelligence," but just the testers' guesses about what questions would assess intelligence; that the research of the IQ developers did not show heredity to be the main causal factor, nor that education is impotent in the face of durable intelligence quotients; and that if these falsehoods were widely believed, it could do immense harm to individuals and to the enterprise of education.

Lippmann conceded that if the tests were presented simply as a bunch of tasks that seem to correlate with tasks asked of schoolchildren, and that the scores simply represented what the test-takers could do at the time they took the test, then using the test results for initial assignment to homogeneous groups in schools might be a reasonable procedure, or, at least, better than other alternatives. But he feared that the "more prominent testers have committed themselves to a dogma which must lead to just such abuse." He concluded that if "the impression takes root that these tests really measure intelligence, that they constitute a sort of last judgment on the child's capacity, that they reveal 'scientifically' his predestined ability, then it would be a thousand times better if all the intelligence testers and all their questionnaires were sunk without warning in the Sargasso Sea."²⁷

It is significant that Lippmann accompanied his forthright criticism with the belief that if the tests were recognized simply as sets of tasks, not indicators of innate ability, it would be reasonable to use them for purposes of classification. The pervasive appeal of the efficiency ideal in large-scale organizations is illustrated by one of Lippmann's own books, *Drift and Mastery* (1914). In that work Lippmann chronicled a transition in America from a "nation of villagers," to a nation of big organizations. The nation of villagers produced populist politicians like William Jennings Bryan, defending a localist culture that clouded one's capacity to understand the twentieth century. In the nation of large organization, rational management and science would prevail. "Rightly understood," wrote Lippmann, "science is the culture under which people can live forward in the midst of complexity, and treat life not as something given but as

something to be shaped The scientific spirit is the discipline of democracy, the escape from drift, the outlook of a free man.”²⁸ Lippmann, one of the era’s most thoughtful and influential liberals, rejected the IQ testers’ claims precisely because they were unscientific and because they brought the threat of gravely undemocratic school practices. But he did not reject the use of tests for classification in schools, nor the ethos of science in the service of social engineering that lay behind it.

Struggles over hereditarian science

Despite Progressive Era reformers who believed in the power of changed environments, this was also the era of Jim Crow laws and of the Supreme Court’s decision in *Plessy v. Ferguson* decision of 1896, in which the Supreme Court of the United States said that state legislatures should enjoy “a large discretion” to “act with reference to the established usages, customs and traditions of the people” in deciding the reasonableness of laws requiring segregated railroad cars or schools. It was the era when America’s imperial ventures were justified in part by the White Man’s Destiny, and when the august arbiter of knowledge, the *Encyclopedia Britannica* flatly proclaimed that “mentally the negro is inferior to the white.” Like racism, hereditarian views had staying power. Expressing them through the concept of IQ was common well into the twentieth century. In the 1930s the famous education professor Ellwood Cubberley said that although some educators had a hard time accepting “biological inequality” with regard to “mental capacity,” they eventually came to see that “nurture and environment . . . cannot to any material extent overcome the handicap of poor heredity.”²⁹

Part of the staying power of such views is that heredity is not irrelevant to mental abilities. It seems, for example, that after controlling as best one can for environment and life experiences, women are better, on average, than men on some math skills and worse on others. Careful geneticists often study such matters, emphasizing that the results are averages, around which there is much intragroup variation, that heredity is only part of the causal nexus of mental abilities, and that scientists should focus on particular skills, not on a general, summary measure of intelligence.³⁰

Many of the leading developers of achievement tests brought to that enterprise the hereditarian concerns of their involvement in the intelligence testing effort. James

Cattell's concern about "race suicide" flowered into diatribes against women's higher education. If Anglo-Saxon women were out of the family, studying, they would be shirking their duty to reproduce. Furthermore, their capacity for higher education was not equal to that of men, and it damaged their health. Cattell supported the ban on women graduate students at Columbia until his colleagues overturned that policy in 1900. And in 1909 he took the cause to the general public in the pages of *Popular Science Monthly*. Women are harmed by advanced education, he wrote, estimating that "to the average cost of each girl's education through high school must be added one unborn child." Worst of all was the "continuously increasing" number of women teachers. These people "subvert both the school and the family."³¹

Heredity, IQ and meritocracy kept bumping into each other. Across 120th Street, at Teachers College, Columbia, Cattell's protégé Edward Thorndike turned from intelligence testing to the development of achievement tests, but he continued to believe that "the world will get better treatment by trusting its fortunes to its 95—99 percentile intelligences than it would get by itself. The argument for democracy is not that it gives power to all men without distinction, but that it gives greater freedom for ability and character to attain power," neatly tucking in the idea that intelligence and character are correlated.³²

However, some substantial progress was made against this Victorian juggernaut all dressed up in the finery of science, particularly on the issue of sex differences. Chicago welcomed women graduate students. There Helen Thompson (later Woolley) studied with John Dewey. In her dissertation (1900) she reviewed the literature on sex differences, exposing inconsistencies and contradictions. Her research, based on physical and perceptual tests of fifty University of Chicago undergraduates, cast further doubt, finding surprisingly random or minor sex-related differences.

Later Woolley collaborated with Leta Stetter Hollingworth in preparing reviews for the *Psychological Bulletin* in the 1910s, expressing an emerging consensus against sex differences in perceptual and mental abilities and debunking various generalizations like differential cranial capacity. Leta Stetter's husband, Henry Hollingworth, was a graduate student of James Cattell. Henry Hollingworth was a feminist, but apparently he disagreed with Cattell without alienating him. Upon receiving his Ph.D., he was

appointed to the psychology staff at Barnard; and when his wife was refused a fellowship in the psychology department at Columbia, he supported her studies, and she got a Ph.D. with Thorndike at Teachers College. Historian Rosalind Rosenberg concludes, somewhat optimistically, that due to the work of researchers like Woolley and Hollingsworth, “by 1920 American psychologists had buried the doctrine of female uniqueness propounded by their Victorian mentors.” Of course, many barriers remained.³³

Indeed, the hereditarian emphasis of Darwinian theory applied to social issues did not succumb quickly or completely. The social Darwinism of Herbert Spencer in England and William Graham Sumner in the United States still held considerable sway in the 1920s. This nativist interpretation of cultural superiority persisted through the restriction of immigration after World War I, and eugenics was still holding its own in the 1920s. Thereafter its empirical exaggerations, the second thoughts by some of its supporters, and much criticism by scientists, pushed eugenics into refuge in the popular hinterland. Nonetheless, as the history of psychology in the twentieth century demonstrates, the debates about heredity and intelligence rise like the phoenix periodically.³⁴ Given the important role that intelligence testing has played in histories of assessment, including this essay, it should be emphasized that IQ testing eventually faded from the school scene, while achievement testing became dominant; and the emphasis on the hereditarian side of mental competencies has also receded except in some periodic debates.

The intellectual world of the efficiency experts

The tool kit of the “administrative progressives” included more than tests and hereditarian theories. By the 1920s, educational leaders at the new, expanding schools of education had created an intellectual structure of psychology, curriculum theory, and administrative principles that worked harmoniously to govern most of America’s schools. David Snedden of Columbia Teachers College devised elaborate inventories and dissections of curriculum that broke down subjects into the tiniest facts or operations, (which he called “peths,” from the Welsh word for a tiny thing). Peths could then be related to “strands” of life activities, which would then be assigned a certain number of 60-hour “lotments,” of classroom work, so that knowledge could be packaged in different curricula for students of different abilities, headed for different roles.³⁵

Franklin Bobbitt, a professor of educational administration at the University of Chicago, provided the central metaphor for the urban school system: “Education is a shaping process as much as the manufacture of steel rails; the personality is to be shaped and fashioned into desirable forms It is possible to set up definite standards for the various educational products. The ability to add at a speed of 65 combinations per minute, with an accuracy of 94 per cent is as definite a specification as can be set up for any aspect of the work of the steel plant.” Bobbitt admired industrial production, he believed that everything can be quantified, and he promised in this essay that if business leaders told the schools what product they want, schools could produce it.³⁶

Thorndike not only developed dozens of tests but became the leading theorist of Connectionism, an early form of behaviorist psychology. It was the perfect accompaniment for standardized tests and Snedden-driven curriculum theory. In Connectionism, learning is the formation of a bond or an association between a stimulus and a response. Bonds are formed through two “laws”: Exercise (practice) and Effect (reward). Since rote learning had been criticized for decades, Thorndike understandably placed particular emphasis on the rewards or punishments that fleshed out the Law of Effect. Bonds are reinforced when accompanied by “a satisfying state of affairs,” and weakened when followed by an “annoying state of affairs.” Of course, the bonds that were formed were akin to “The capital of New York is Albany,” and the danger was that with so many facts in the world, the curriculum would be flooded with them and not with thinking about how the facts got that way, or how they related to each other, or what their significance was.³⁷

The same institutions at which these ideas were generated—Teachers College-Columbia, Chicago, Iowa, Stanford and others—also became training grounds for thousands of young researchers and administrators who would become principals and superintendents in large cities, work in state education offices, and promote further research and policy advocacy regarding elementary and secondary education, all of which served to expand testing. The senior professors at these leading schools of education formed networks, such as the “Cleveland Conference,” an annual meeting for fifty or sixty self-appointed members, at which they talked about education reform, administration, and, presumably, the placement of younger rising stars.³⁸

The expansion of achievement testing

Achievement tests had precursors in nineteenth-century written assessments, such as high school admissions tests. Some of these mid-nineteenth-century tests hinted at critical thinking and understanding, as in these examples: “state the causes of the American Revolution,” or “describe the settlement of Pennsylvania.” But most test answers depended upon factual recall from rote learning, like “Who discovered Florida. . . the Pacific Ocean. . . the Hudson River,” “Define: Promissory Note, Bank Discount, Present Worth,” or “state the bounds, divisions, climate, productions, religion, and seven principal cities of Italy.”³⁹

In the early twentieth century, achievement tests underwent a flurry of development, parallel to but soon outpacing the use of IQ tests. The object of such measurement was not native ability but things learned, put in a comparative context with other students, other schools, and national norms. Spurred by the growing scale of urban school systems, by the burgeoning proportion of immigrant children in the schools, by the ethos of efficiency and measurement promoted by corporate capitalism, researchers at professional education schools developed tests to measure achievement in the basic school subjects.

In 1908, a student of Thorndike named Cliff Stone published an arithmetic reasoning test, credited as the first standardized achievement test. In 1910 his mentor Thorndike published a handwriting scale, which provided models with which teachers could compare students’ handwriting. In 1913 another Teachers College product appeared, the Buckingham Spelling Scale, with words listed in order of difficulty determined by field-testing. Leonard Ayres, a competitor from the Russell Sage Foundation, published the “Three-Slant” handwriting scale in 1912 and then, in 1915, a spelling scale that listed 1,000 rudimentary words in order of frequency, from “the” to “wreck.” This was the seedtime for achievement tests. The subjects covered rapidly multiplied, including Daniel Starch’s grammatical scale, punctuation scale, grammar test, and Latin vocabulary test; Thorndike’s “visual vocabulary,” and a scale for “the merit of drawings by pupils,” plus his scale of ability to understand sentences; and Walter Monroe’s standardized silent-reading tests, among many others, all before 1920.⁴⁰

By the 1920s most achievement tests used a multiple-choice format, which eliminated any ambiguity about the correct answer. Advocates emphasized that this made the scoring not only faster but also completely “objective.” Arthur Otis is credited with providing the model, for use in the Army Alpha intelligence test in 1917. This feature soon became widespread in achievement tests.⁴¹

This burst of commercially available standardized achievement tests paralleled the IQ testing movement and was supported by the same intellectual and institutional framework, achievement tests soon surpassed intelligence tests in market share and in school use. Ultimately they had a much greater influence on school practice.

There were several reasons for this greater impact. First, intelligence tests were based on the premise that a single score on a single test represented a child’s “general” intelligence; thus one test would suffice. But achievement tests ranged across the different subjects of the curriculum, and soon the testing industry had achievement tests for dozens of subjects. Thus they constituted a larger part of the market. Second, although IQ tests were useful for classifying and assigning students to tracks (if you believed in the premises), they were not useful in measuring the results of teaching and learning. Districts could use achievement test results to judge whether their curriculum was aligned to the implied curricula of national tests and how effective their instruction was. Third, intelligence tests advocates made the baldest claims about the heritability of mental ability and its permanence across an individual’s life course. These arguments attracted the most strenuous criticism, on both empirical and philosophical grounds. Despite the persistence of hereditarian views of intelligence, opposition to the use of IQ tests for consequential school decisions grew from decade to decade up through the 1960s, by which time few elementary or secondary schools used it for placement.

The proliferation of achievement tests continued in the 1920s, with major development centers at Stanford (led by Terman), Teachers College, Columbia (led by Thorndike), and Iowa (led by E. F. Lindquist). Mathematics, history, and other subjects joined the list, along with attempts to measure character, personality, interests, and preferences. Bibliographies and surveys catalogued hundreds of such tests from the 1920s to the 1940s. In 1954 the *Fourth Mental Measurements Yearbook* described 4,417 tests, of which about one-fourth were intelligence tests. Corporations grew up around the

testing business. Test producers provided scoring aides, first through such mechanical devices as stencils, then in the 1940s, machine scoring from IBM.⁴²

Professional educators also spread the word, through books about testing, conferences, articles and yearbooks. Two developments generated energy for the testing enthusiasm: research bureaus and school surveys. City school districts and state education agencies established bureaus of research, often staffed by recent graduate of doctoral programs in the major test-producing universities. School surveys were comprehensive studies of single school systems by a team of expert outsiders who examined the organization, finances, physical plant, supervisory relations, teachers' credentials, classroom methods, curriculum, and the achievement test score results of students. The results were presented to the school board and staff; sometimes, depending upon the contract, the reports were public.⁴³

Not all the administrative progressives wholeheartedly supported the use of test scores to rate educational success. Historian Tracy Steffes, who has studied the development of school systems extensively, concluded that some of the discussions were quite nuanced about the limitations of measurement, the possibility that it might crowd out less quantifiable educational objectives, and other possible unintended consequences. One critic warned, "There is a real danger that efficiency, having become a fetish, may be pursued purely for efficiency's sake." Nonetheless, most administrators, on balance, welcomed tests and surveys as scientifically valid ways to improve education. Leonard Ayres reported in 1912 that 216 cities had established systems of student records to keep track of their progress, an increase from only 29 three years earlier. Responding to the charge that "advocates of the scientific method aim to reduce all work in education to the dead level of uniform precision," Ayres said that, on the contrary, the aim was to provide the appropriate education for each individual. This use of test scores to rate the effectiveness of schools was an impulse of the early twentieth century system-builders; it seems to have become submerged in the 1930s, to be revived again in quite robust form, after 1965, as we shall see.⁴⁴

Higher education and testing

Colleges were drawn in to the multiple-choice testing regime through standardized entrance examinations. In the late nineteenth and early twentieth centuries, many colleges had their own entrance examinations. Some sent faculty members out to certify high school programs in an effort to regulate quality. The College Board had been established to develop standards and examinations for college entrance, and they began administering written exams in various subject areas in 1916. In addition to these tests they developed in 1926 a multiple-choice test designed to assess college readiness. This test was a direct descendant of the Army Alpha intelligence test from World War I. Carl Brigham, who worked for Yerkes during the military testing experiment, developed this Scholastic Aptitude Test, essentially a more difficult version of the Alpha. Brigham was a professor at Princeton and a eugenics enthusiast. Among his friends were Madison Grant and other theorists of racial hierarchy. In his *Study of American Intelligence* (1923), Brigham ranked three white races, the Nordic, the Alpine, and the Mediterranean, in descending order of average intelligence. He concluded that because of the shifting immigration patterns to the United States, the average American IQ was falling.⁴⁵

Despite cautionary statements from the College Board about the Scholastic Aptitude Test's accuracy in predicting later performance, the SATs were clearly aimed for a market seeking exactly that. As Nicholas Lemann points out, Brigham later had a total change of mind and recanted his hereditarian beliefs. In a book significantly titled, *A Study of Error* (1932), he argued that intelligence tests did not measure something hereditary or properly associated with ethnicity. He tried to distance the SAT tests from the IQ testing enterprise, and he opposed the formation of an organization to hasten the spread of the SAT to the whole field of college admissions. He managed to slow it down, but the machinery was in motion to market the SAT, and after Brigham's death in 1943, the mainstay of opposition was gone. The SAT became a major project of the College Board and the new Educational Testing Service, established in 1947 with the support of Devereaux Josephs of the Carnegie Foundation, Henry Chauncey of the College Board, and James Conant, President of Harvard University. The iconic importance that SAT scores assumed in admission to elite colleges, and the notion that it was an "aptitude" test, not an achievement test, formed a shadow of IQ testing over the SAT. Conant

actually thought that the SAT's alleged independence from school learning was a virtue since rich people got better schooling for their kids in elementary and secondary years, and the SAT would tend to level the playing field. Generations of affluent people buying test preparation to improve their children's "aptitude" would prove the naiveté of calling the SAT a measure of "aptitude." Nonetheless, the SAT and its eventual competitor in the Midwest, the American College Test, became fixtures in higher education admissions.⁴⁶ This is not to say that there is no merit to the idea of a meritocracy. Many students from non-affluent backgrounds have benefitted from strong performance on the SATs. But the test came to have an IQ-like aura, the magic bullet of college admissions. To the extent that the test in fact assesses the verbal and mathematical knowledge that is privileged in privileged families, it was a flawed vision.

The Scholastic Aptitude Tests became very influential in college admissions. They occupied a ground between intelligence tests and achievement tests. They retained their prominence in college admissions longer than IQ tests, but their claim to measure "aptitude," rather than learned cultural content, increasingly came under fire on the basis that they were class and racially biased. Despite efforts by Educational Testing Service to address and diminish bias, many colleges in recent decades (830 four-year colleges by 2010) have de-emphasized the importance of the SAT's in their admission process.⁴⁷

Prelude to reform: The 1940s and 1950's

World War II left the United States as a major world power, with a huge military machine, a much larger and more active federal government, and an expanding education system. High school attendance among 14—17-year-olds, which had been 32 percent in 1920, was 73 percent by 1940.⁴⁸ The percentage of students completing high school and receiving diplomas rose apace. By 1950 high school graduation was the modal experience of American youth. The logic of a differentiated high school curriculum seemed ever more obvious. Higher education expanded, especially in the community college sector. With increasing federal help, research expanded as well.

The rationale for testing and placement was longstanding and widely accepted. Of course, there was debate and criticism. In an essay on "Five Decades of Public Controversy Over Mental Testing" Lee Cronbach argued that the reception of testing

critiques depended upon the zeitgeist, the conditions and predominant beliefs of the time. Thus, Allison Davis's criticism in the late 1940s, that educational testing was biased against working-class students, got little traction. His complaint that working-class kids had little exposure to the kind of vocabulary that it took to score well on school tests did not have much of a constituency; the times were not right.⁴⁹

By the late 1950s the international context of education reform became salient again. In the late 1950s educational reform in the United States was energized by the successful launch of the Soviet satellite *Sputnik*, which persuaded enough Congress members and a reluctant President Eisenhower to support federal aid to elementary and secondary education, embodied in the National Defense Education Act of 1958. The evidence for this crisis was not test scores; that criterion had not been established yet. Advocates for NDEA argued that Soviet superiority was obvious from the technological achievement in space and that all Soviet children were exposed to a challenging school curriculum full of advanced math and science. These views turned out to be exaggerated, but they carried the day, with testimony in Congress by such luminaries as Admiral Hyman Rickover of the U. S. Navy. The NDEA acknowledged the central role of testing in U.S. schools by providing states with funds to help local school districts increase testing and guidance, with an emphasis on finding the most talented children in America and encouraging them to become scientists. While the NDEA had a Cold War origin, it also resonated with the agenda of the administrative progressives of the early twentieth century: emphasis on testing, guidance counseling, the identification of talent, and the differentiation of curriculum.⁵⁰

By the end of the 1950s testing had even more pervasive in educators' professional culture and in the daily life of their schools, than it had been before the Second World War. Research universities and publishers continued to produce new tests, and James Conant's book on *The American High School Today* laid out plans for the "comprehensive American high school," which proclaimed a meritocracy of students who were chosen for more difficult or less difficult courses on the basis of academic ability, but not kept strictly to a single "track" in all subjects. Conant was trying to articulate the virtues of those already existing schools with the basic characteristics of the "comprehensive high school." This plan required expanded testing and counseling roles

for the school personnel. This became a common theme at professional meetings and was reinforced by the National Defense Education Act, which subsidized the development of testing and guidance.⁵¹

Confidence in the “meritocracy” was soon to be questioned, however, by the increasing civil rights concerns of the 1960s and beyond. Along with concern about cultural bias of tests and their inequitable impact on different groups, there was a major revival of the idea that program evaluation and accountability should focus on students’ results on achievement tests.

Testing and Civil Rights

The activist 1960s phase of the long civil rights movement brought much greater concern for equity to the field of testing. This meant criticism of some tests but also a contrasting pressure to do more targeted testing of different groups’ performance in order to determine whether schools were serving students well or whether specific reformed practices were working. Critiques about the disparate impact of many tests on students of different economic classes or different race or ethnic groups had substantial impact, leading to a further decrease in the use of IQ tests for placement, efforts to reduce or eliminate cultural bias in other tests, such as the SAT, and restrictions on the uses of various tests in educational settings, emanating both from courts and legislatures.

The focus shifted from equal opportunity to equal outcomes. A notable early step was taken in employment law. The Civil Rights Act of 1964 included in its Title VII provisions for fair employment practices; interpretations of Title VII in guidelines produced by the Equal Employment Opportunity Commission (EEOC) in subsequent years revealed a shift in the goal, from equal treatment to equal outcomes. The Supreme Court endorsed this interpretation in a 1971 case, *Griggs v. Duke Power Co.*, which declared that EEOC’s rulings had the force of law and that the goal was to prevent “disparate impact,” that is, a pattern in which the average scores among test takers from one racial group are consistently lower than those from another group, which was increasingly disallowed by courts, especially when the test content could not be proved to be directly relevant to the job.⁵²

These guidelines and decisions spilled over into uses of tests in the education sector. In the appellate court decision, *Singleton v. Jackson Municipal Separate School System* (1969), the justices prohibited the use of tests to re-segregate black students through tracking. In a parallel case dealing with special education, the Supreme Court declared in *Larry P. v. Riles* (1972) the San Francisco schools could not use IQ tests for placement of black students in special classes for the mildly mentally retarded. The tests, said the Court, were not valid for this use. Thus began the demand that tests be scientifically validated for the uses to which they were put. In the Education of All Handicapped Children Act three years later, often referred to be its Public Law number 94-142, the Congress adopted and spelled out such validation requirements.⁵³

The civil rights movement, then, led to many restrictions on the uses of tests. Nonetheless, we know from studies of the testing industry that the purchase and use of tests continued to rise sharply during these and subsequent decades. One reason is that although civil rights advocates initiated critiques that led to restrictions on the uses of tests, they also contributed to an increased *emphasis* on tests because of concerns about the highly publicized gap in test scores between students of color and white students, and between poor students and more affluent students. The achievement gap, then, led to demands for less discriminatory tests but also for more equal results in the education of all children. Reducing the achievement gap became an education goal from the 1960s to the present, particularly on the part of the federal government. In many settings, achievement scores by race-ethnic group became the leading factor in people's judgments about whether integration was working successfully; indeed, in some cases, closing the achievement gap became a substitute for integration.⁵⁴

This focus on test scores was reinforced by several factors, several of them relating to accountability, not of students but of teachers, schools, and school districts. This function of tests was nascent in some earlier practices, like the school surveys of the early twentieth century, but the 1960s was a watershed from which many forms of accountability flowed.

The federal government and the birth of national assessment

When Francis Keppel arrived in Washington as the Commissioner of Education in the Kennedy administration, he was frustrated to discover that the Office of Education seemed to have no data about what was learned in schools. In October, 1963 he wrote a friend that so far the OE “seems to have reported about almost everything except what children learn.” He said, “with the understanding that it is in confidence,” he hoped to build a “reporting system on quality of education in a form useful to local schools.”⁵⁵ In a 1968 interview Keppel said that the Office needed not just numbers of teachers and schools—the typical OE fare—but data on “what learners learned,” on the “quality of institutions,” and on the evaluation of new programs. He realized that such an initiative would “lead to outcries about Federal control over education,” so he went to his friend John Gardner, president of the Carnegie Corporation, and asked him if he would explore the issue and come up with a plan under which “you are not testing every kid . . . but get a sample which will give us some benchmarks.” Gardner agreed.⁵⁶

Gardner enlisted the help of psychologist Ralph Tyler, founding director of the Center for Advanced Study in the Behavioral Sciences at Stanford. Carnegie funded two conferences. Then in August 1964 they established a standing Exploratory Committee on the Assessment of Progress in Education. Over the next few years the Exploratory Committee developed a model for the National Assessment of Educational Progress. It featured matrix sampling (no individual would answer all the items), and was criterion referenced (the nation’s scores would be compared to learning objectives regarding what should be known in a subject, determined by committees of experts and other citizens), with a mixture of multiple-choice and short answer items.⁵⁷

Keppel was right to think that the venture was controversial. As word of this possibility circulated among school administrators, alarm bells went off. Writing in the September 1965 issue of *Phi Delta Kappan*, Harold Hand, a University of Illinois education professor, criticized the secrecy of the project and reasserted the criticisms he had recently made at a meeting of the Association for Supervision and Curriculum Development. Hand argued that the national assessment program would make schools compete with each other. Rating schools against each other would drive good teachers away from teaching the lower achieving students, a “hideous consequence” for equal

educational opportunity. Furthermore, national testing would lock districts into a single curriculum, stifle local experimentation, and undermine local control. In the same issue, Ralph Tyler made the first public report on the status of the project.⁵⁸

The fact that the assessment had never been intended to produce scores for individual students or schools only underlined the secrecy about which Hand complained. In the same issue of the *Phi Delta Kappan*, the first public report on the Exploratory Committee's work, Ralph Tyler assured his readers that "no individual pupil, classroom, or school shows up at all" in the planned assessments. It was designed only to show what all the students in the nation, at a given age, can do in math, reading and numerous other subjects, on average, and to give percentages for what percent of students were at various levels of successful answers. He said the development of the test was moving slowly and carefully, involving panels of teachers and other educators, and would be ready soon for trial runs in selected districts. The committee had contracted with various research organizations (American Institutes for Research, Educational Testing Service, and the Psychological Corporation) to develop full-blown tests in different subjects.⁵⁹

In August of 1965, one month after passage of the big Elementary and Secondary Education Act (ESEA), Keppel answered a query from White House domestic chief Joseph Califano about his priorities. At the top of his list was to "assess where we are in education." Evidence was "woefully lacking," said Keppel. He admitted that it involved "delicate problems." "Many educators, perhaps even the majority, are dubious about any kind of national testing" because it would suppress "imaginative teaching," make unfair comparisons "at the expense of the disadvantaged," and lead to "rigid" control of curriculum by the federal government. Then he described the Carnegie exploratory work, which would lead to ways of assessing progress without making unfair comparisons, "measured in a fashion broad enough to give local freedom in managing the schools."⁶⁰

Further concerns made their way from Congress to the executive branch. In September 1965 Keppel wrote in response to Senator Roman Hruska that the assessment would only be a sample of schools and that no scores for individuals were possible. The purpose was to judge whether education in the U.S. was improving. Although he had not funded it, Keppel said, he had encouraged it, because he assumed that Congress would expect "firm evaluative data concerning the educational programs being supported with

Federal funds.” He enclosed Tyler’s recent article and a statement by the Exploratory Committee laying out these policies and plans. In October, Under Secretary of Health, Education and Labor (HEW) Wilbur Cohen sent a similar letter to Senate minority leader Everett Dirksen, who had inquired on behalf of a distressed constituent. He said that the Office of Education was not officially involved but hoped that the project would help the Office of Education to report responsibly on federal programs in education. To do this “merely by describing how the funds have been spent,” said Cohen, was insufficient. “Real evaluation requires measure of how learning has increased as a result of programs.” Cohen said that HEW stood firmly behind the principle of local control. He enclosed the Exploratory Committee’s statement to be sure the constituent realized that not all students were going to be tested.⁶¹

Despite these clarifications, national testing remained a flashpoint for those who were concerned about local control. This had been, of course, a perennial theme in American school governance, the sacredness of local and state control. It had been a huge factor to be overcome in the passage of the National Defense Education Act of 1958, overcome in part because of the visceral issues involved in Russia’s launching of the Sputnik satellite. The effective coalition between conservative Republicans and segregationist Democrats had frustrated federal aid advocates during the Kennedy administration, calling upon the ready reservoir of popular belief in local control.⁶²

In October 1965 *U. S. News and World Report*, the more conservative of the three newsweeklies, mentioned the assessment plan as part of a pattern of federal interference in education. In an article that was mainly about Keppel’s ill-fated attempt to pressure Chicago into desegregation, the *U.S. News* editorial added that other cities were scheduled for such investigations, and rumors flew that the Civil Rights Act was creating federal pressures on textbook publishers to use more examples of minority group members in their texts. Furthermore, a Congressionally-mandated questionnaire connected with the ESEA was resisted by eight cities because it asked teachers and students to answer questions like, “If you could have anyone you wanted for your close friends, how many would be white?” And finally, said the editors, a second “testing program,” the Carnegie-sponsored national assessment, “worries those who want the Federal Government to keep its education activities to a minimum.”⁶³

During the coming year Keppel continued his attempt to get across the nature of the national assessment. In February 1966, he spoke at the annual meeting of the American Association of School Administrators in Atlantic City. On a panel, Keppel first listened to Cincinnati's Superintendent of Schools criticize the federal government for expanding its programs without adequate funding or explanations. Then, in his presentation, Keppel stuck to his agenda. He said that the federal government needed to get "better information on the condition of education within the states." He endorsed the Carnegie effort to develop a national assessment, which, he said, was a "fact-finding, rather than a testing program."⁶⁴

With the impetus of the big elementary and secondary bill behind him, Keppel persuaded the Congress to go along with the assessment, now officially labeled the National Assessment of Educational Progress (NAEP). In 1968 the Office of Education gave a grant to the exploratory committee to implement the development and inauguration of NAEP, which was implemented in 1969—1970, under a contract to the Education Commission of the States, a recently formed nonprofit organization headquartered in Denver. In 1982, in a competitive process, the NAEP contract went to the Educational Testing Service (ETS) in Princeton. Over the years, compromises and changes were made. NAEP had begun by resolving to combine multiple-choice questions with short answer and essay responses, but gradually they shifted predominantly multiple-choice items. A second change that involved trade-offs was that, under ETS, Item Response Theory (IRT) was developed and applied to the psychometrics of the assessment. IRT was technically superior and efficient, but it reduced the items at the very top of the difficulty range and at the very bottom, the emphasis being on items that would discriminate between assessment subjects, more than inventorying a lot of things that most students could do, or that most could not do. In 1988 Congress created an oversight board for NAEP, called the National Assessment Governing Board, and authorized the release of NAEP scores by state, a process deeply studied by both the National Academy of Education and the National Research Council. Many assessment experts resisted any further disaggregation of NAEP scores and especially argued that using NAEP for individual student scores would compromise its value as a monitoring device of learning in the states and in the nation.⁶⁵

Accountability as a major goal of assessment

Parallel to these efforts to establish a national assessment of educational progress, at least three other developments pushed the idea of basing policy decisions upon measurements of student learning: Robert Kennedy's amendment to the Elementary and Secondary Education Act (ESEA) requiring testing for accountability; the vogue in the federal government to introduce cost-benefit analysis into policy evaluation; and the signal influence of James Coleman's 1966 report using achievement scores to assess the influence of school resources and family variables on student learning.

Robert Kennedy's accountability amendment

In the spring of 1965 the House of Representatives passed the Elementary and Secondary Education Act," under considerable pressure from the White House to move swiftly. Then there was even more pressure for the Senate to pass the House version without amendment, to avoid delay and debate in a conference committee. Senate Wayne Morse of Oregon, the bill's manager, attempted to implement this plan. But a feisty new Senator from New York, Robert Kennedy, raised the issue of whether ESEA would require districts to produce any evidence that its programs, particularly Title I (compensatory education), worked effectively. This issue did not engage as many Congress members as the other issues, but the debate was charged with electricity because when Robert Kennedy was Attorney General in the Kennedy Administration, he had been a thorn in Vice President Johnson's side, and there was great tension in the relationship. Even RFK's friends thought that his constant public criticism of Vice President Johnson's performance as chair of the Committee on Equal Employment Opportunity had been unfair and shabby.⁶⁶

Now, as Senator Kennedy squared off for battle with the Johnson administration, he drew on his experience as a member of President Kennedy's Task Force on Juvenile Delinquency. Various members of that task force had concluded that American public schools were dysfunctional for poor and minority children and were thus part of the problem, not part of the solution. That put Kennedy on a collision course with the President, whose strategy was to pass the education bill now and fix its flaws later. According to columnist Joseph Alsop, Kennedy consulted widely about urban education

in preparation for this effort. Through his education aide Adam Walinsky, Kennedy conferred with James Allen, commissioner of education in New York, President John Fischer of Teachers College, Columbia, and other experts, as well as with education advocates in the House of Representatives.⁶⁷

Kennedy was on the Senate's General Education subcommittee that held hearings on ESEA. When Frank Keppel appeared as a witness, Kennedy asked him whether schools were not part of the problem in educating the poor. Keppel replied, "I am sorry to say that is true." Indeed, a year earlier Keppel had told a national meeting of school administrators that federal aid alone would not be enough to connect education and the amelioration of poverty. It would take changes in local attitudes and local practices. "We are simply not reaching hundreds of thousands of children who are now in our schools." We relegate children of the slums to a "lower order" by perpetuating terms like "culturally deprived" and use these terms "as alibis for our failure" to educate these children. "These children need the most skilled of teachers . . . the least crowded of classrooms . . . the best of educational opportunity," not the opposite. They need "imaginative change from ways of teaching which do not work." They need books that "recognize their existence in our society." Rising to fervent oratory, the Commissioner had concluded, "When our judgment comes, let it be said that we gave the excellence that was in us to these schools of poverty."⁶⁸

With equal fervor, Robert Kennedy now challenged Keppel. If we pour money into this school system, he said, "which itself creates this problem, . . . are we not just in fact wasting the money of the Federal Government?" As sympathetic as he was to this point of view, Commissioner Keppel could not agree with that conclusion publicly. He mustered instead a weak optimism, replying that school people were undergoing a "rapid change in attitude." Columnists Evans and Novak chortled, "The soft underbelly of President Johnson's much-praised program of aid to education is being attacked by none other than the Senate's most famous freshman."⁶⁹

Nonetheless, Keppel supported Kennedy's desire for accountability. As we have seen, he was determined to collect more information about student learning through a national assessment. But in the case of the Elementary and Secondary Education Act, he was also sure that if an amendment requiring federal testing for all Title I schools were

added to the legislation, it would doom it. A deal was struck. Keppel drafted an amendment that would require districts to devise their own tests and to report the results periodically to the state education agencies. In return, Kennedy agreed to the administration's request that the amendment be introduced, not by Senator Kennedy but in the House of Representatives, so that the Senate could simply pass the House version without changes, avoiding the delay of a conference committee. Representative John Brademas of Indiana, who supported the ESEA but shared Kennedy's skepticism about whether Title I would work, introduced the accountability amendment in the House, and it became part of the bill.⁷⁰ This assessment provision was weakly drafted and weakly enforced. In reality evaluations of Title I that included test scores were mainly performed by research organizations or were done by the government through samples of districts.

A budgeting brainchild from the Pentagon:

Along with the faint beginnings of accountability requirements in Robert Kennedy's amendment to the Elementary and Secondary Education Act, a second development in the mid-1960s advocated accountability for results more generally across the federal government. Secretary of Defense Robert McNamara had instituted a budget mechanism called "Planning—Programming—Budgeting System" (PPBS). Facing tough budgets and rising costs of the War in Vietnam, President Johnson adopted it for all government agencies in the summer of 1965. The essence of PPBS was to require departments to arrange for independent budget and performance analyses for their programs, so that at a higher level, budget makers would not have to rely solely upon the requests and rationales of the program managers. Because McNamara was deemed to have taken effective control of the massive Defense Department and its budget process, the White House concluded that PPBS was the reason for his success. Joseph Califano, the new domestic policy chief, had come to the White House only a few weeks before President Johnson announced the required adoption of PPBS. Califano, who had served the previous year as a Special Assistant to Secretary McNamara, was a firm supporter of PPBS.

The decision to require PPBS from all departments had many critics. Its implementation was tentative and confused. It ran as a parallel budgeting exercise,

alongside the existing budget-setting mechanisms, and it remained on the margins. Five years later, President Nixon cancelled the use of PPBS. Nonetheless, it had some lasting effect by promoting the notion that agencies needed to evaluate programs and base their budget requests on documented cost-benefit analysis. In an interview, Emerson Elliott, the principle examiner of education programs for the Bureau of the Budget, recalled the flurry of activity and White House pressure connected to PPBS. The input-output emphasis was not solely the brainchild of Robert McNamara; various enthusiasts pointed out that it was part of a larger, more gradual shift to focus on outcomes.⁷¹

Although that shift was widely embraced by policy analysts and critics of programs, it encountered much opposition at the ground level. Laurence Lynn, who chronicled the demise of PPBS, provides an epitaph for the program: “The birth of PPBS was economic common sense. It died as political common sense.”⁷² Agencies did not want to be responsible for the measured consequences of their programs, and they had the power to resist. Still, the emphasis on outcomes became a central focus of policy analysis, and it ultimately had some effects on policy-making and legislation.

The Coleman Report

A third influence that reinforced the emphasis on student outcomes was James Coleman’s study, *Equality of Educational Opportunity*. Congress mandated this research in the Civil Rights Act of 1964; its purpose was to explore inequality in the educational opportunities of black and white students. Commissioner Keppel’s new in-house statistical expert, Alexander Mood, insisted that the researchers look at educational outcomes, not just resources of schools. Mood knew sociologist James Coleman’s work in the statistical analysis of large-scale surveys. Coleman, then at Johns Hopkins University, accepted Mood’s invitation to conduct the study, and he presented the full report to the government within the scant two-year period allowed.⁷³

Coleman’s massive, statistical study of resources and outcomes in schooling was a landmark in methodology and a potential bombshell in its implications for education policy. The data for the report included background characteristics of students: family variables, race, and region, school resources ranging from quality of facilities (for example, existence and size of libraries, ample provision of texts), as well as the training

of teachers. The research also included a wide-ranging questionnaire about family characteristics, student attitudes, and data on teacher quality. Finally, the Coleman research gathered achievement scores. Almost 600,000 students participated, from 4,000 schools, both elementary and secondary.

Complex statistical procedures, beyond the understanding of most readers, suffused the 700-page report. When the major findings finally emerged to public view, they included the following: differences between the resources of schools attended by white and black students were far less than anticipated and had a weaker relationship to achievement scores than expected. The more salient variables associated with test scores were: the preparation, verbal abilities, and confidence of teachers; the student's family background, including class and race; the social class of a student's classmates; and students' convictions that they could control their fate.⁷⁴

The Office of Education did not know what to do with the Coleman Report. Not only was it difficult to comprehend, but its apparent implications were unwelcome to liberal educators and school administrators alike. It asserted that most school variables had little efficacy in improving achievement scores. This threatened not only support for the new Elementary and Secondary School Act, but the longstanding belief that schools were the agent of opportunity for all students. It suggested that spending resources on schools was less important than previously believed. The report also seemed to suggest that the impact of classmates on a student's achievement was more driven by class than race, which raised questions about the value of racial desegregation, an active priority of the Office of Education, a priority that was encountering fierce resistance from the South and some skepticism in the North.

Rightly concluding that hardly anyone would be able to read the actual report, Commissioner Howe presided over a process of developing an Office of Education summary. The result was highly influenced by those in the OE who were cringing about the implications of the Coleman Report for school budgets and for integration. Thus they ended up with a vague summary that dodged the most surprising and tough conclusions of the larger report. Commissioner Howe also evaded the more controversial findings in speaking to the press. In the end, the complex but nonetheless startling implications of the

Coleman study had little immediate effect on the Office of Education and its programs, or on the Johnson White House, or on Congress.⁷⁵

However, the Coleman Report had a significant long-term effect upon the education research community by boosting the relationship of quantitative social science to policy analysis. In the short run it prompted a minor industry of criticism, commentary, and further research by social scientists. Several of these critical studies either re-worked the Coleman data or provided new empirical evidence on the relationship of school resources to student learning. Arguing variously that Coleman either had used inadequate statistical techniques or a flawed research design, subsequent studies concluded that school effects were considerably greater than Coleman had estimated. But these results were a long time coming, and were not available to the journalists and government officials trying to make sense of the report at the time.⁷⁶

In any case, the Coleman report dramatized the newfound emphasis on outcomes. In the process, it shocked many people by the possibility that schools were less efficacious than they wanted them to be. The importance of outcomes is a legacy that lives today, however imperfect its results in practice. And the reduced expectation that schools can solve social problems—even though Coleman may have underestimated school effects—is a disquieting refrain in our education policy culture.

The 1970s

The Vietnam War had produced not only tragic loss of life and domestic political turbulence but also budget pressures that curtailed domestic reform programs. With the onset of a Republican administration, the enthusiasm for desegregation was tempered; the White House sought new approaches in education reform. Nonetheless, the civil rights movement fanned out and inspired the already existing momentum of rights movements on behalf of women, language minorities, and children with disabilities, leading to important initiatives and landmarks, notably: a series of Bilingual Education acts, starting in 1968; Title IX of the 1972 Education Amendments, which banned sex discrimination in education programs receiving federal funds; and the Education of All Handicapped Children Act of 1974 (widely known as “P.L. 94-142”), which encouraged maximum feasible participation of children with disabilities in regular classrooms and required that

each special-needs student have an Individual Education Plan. These movements and associated legal developments increased the amount and significance of testing, from the tests used in classifying and placing special education students, to the tests demanded by courts to assess equality of opportunity for students of color.

Beyond Title I's compensatory education programs, the new emphasis on reporting achievement test scores had some ill effects in the 1970s. The minimum competency movement was fueled by unemployment, declining test scores on measures of literacy and math skills, and publicity on the connections between joblessness and low skills. High stakes tests put pressure on students facing the specter of failing a grade or not graduating; but it also had the effect of dumbing down the curriculum by its emphasis on multiple-choice tests and its focus only on minimum standards.⁷⁷ The movement to toughen high school graduation requirements by testing for basic skills was part of a larger, growing movement for accountability of systems and of teachers as well as of students. Like the minimum competency testing programs, this was a state-level movement, partaking of the emphasis on outcomes that had its roots in the 1960s.⁷⁸ Criticism of minimum competency testing did not quash the movement in the 1970s, but in the 1980s it was absorbed or replaced by more comprehensive state-level reforms.

The early 1970s also brought a reminder that the debate about the relative impact of inheritance and environment on mental ability was only sleeping, not dead. Arthur Jensen revived it dramatically with his article "How Much Can We Boost IQ and Academic Achievement," which argued that a very high proportion of IQ was associated with heredity rather than with environment. Citing differing average intelligence scores by race, Jensen expressed doubt about the effectiveness of compensatory education for low-achieving students of color.⁷⁹

A flood of commentary appeared during 1970, and the debate continued with Richard Herrnstein's 1971 article "IQ" in *The Atlantic Monthly*. These and the various responses demonstrate the durability of the debate despite the decline in the use of IQ tests for high stakes in education. The critics argued, first, that the relative weight of heredity and environment is difficult to estimate, and calculations depend upon numerous arbitrary choices of statistical procedures, second, that the either-or nature of such estimates overlooks the important role of interactions between genetic and environmental

factors, and third, that for policy decisions about education, the relative weight is not relevant.

There is no doubt that both heredity and environment affect mental traits, like other traits. However, the meaning and nature of the tests themselves, as well as the statistical methods used to interpret mental ability tests, and the connections made between the findings and policy prescriptions are often questionable and emotionally charged.⁸⁰

The 1980s: A Nation at Risk?

During the 1980s much of the impetus of school reform (and thus attention to achievement testing) moved from the federal government to the states. There was a “push” factor and a “pull” factor. President Reagan eschewed an active federal role in education and wished to devolve the initiative for education policy and reform onto the states; at the same time, more governors were convinced that improved education was critical to their states’ development. Education moved up on the governors’ priority lists. Reform was driven by a shaky economy and prominent journalistic attention to declining SAT scores and international comparisons. The famous Department of Education report, *A Nation at Risk* (1983) trumpeted these anxieties about international economic competition.⁸¹

The first reaction was to toughen high school graduation requirements, raise the standards for admission to teaching education programs and urge tougher local school conditions: more homework, stricter attendance rules, and the like. These tactics proved ineffective, at least in the short run. In the assessment arena, much of the talk was about excellence; simultaneously there was talk about back-to-basics. The governors, their chief state school officers, and education specialists in state legislatures continued forging state-level reform bills for increased salaries, more testing, better training, and tougher standards. In several states, from the late 1970s to the early 1990s, there was a movement to require tougher state oversight, district accountability, and state achievement testing systems.⁸²

President George Herbert Walker Bush, succeeding President Reagan in 1989, showed interest in restoring more federal participation in partnership with the states. A

“summit” of governors was held at Charlottesville, Virginia in 1989, calling for renewed reform efforts. Some governors, notably Bill Clinton in Arkansas, were supporting state education reform programs that would become the basis for a renewed entry of federal action in the 90’s.

By the later 1980s, despite much reform activity, educators were concerned that school reform was sputtering. In the *Phi Delta Kappan*, a monthly magazine widely read by school professionals as well as researchers, Chris Pipher reported in June of 1987 that policy makers in the states believed that the movement was experiencing “its slowdown, its demise, problems of implementation, and lack of results.” Congress was concerned that the advocates for Title I had never produced persuasive hard data about the effectiveness of compensatory education programs. In the reauthorization of the law in 1988, Congress required states receiving funds to define the levels of achievement students should attain and identify schools that were failing to meet those goals. Testing had become not only the spur to reform but also one of the main instruments of reform.⁸³

Psychologist James Popham put forth an argument consistent with educators’ concerns that test results showed a school system with students at too low a level of competence. He advocated measurement-driven instruction. Popham argued that teaching to the test would be a good thing if the multiple-choice tests were well designed, well explained, and “criterion referenced”—that is, the goals were set by the nature of the content desired, not by national scores turned into norms. High-stakes tests would improve learning. Arguing against that view, test expert Lorrie Shepard replied that the data suggested little real improvement where such practices had been attempted, that the kind of testing Popham advocated would inherently lead to teaching to the test in the worst sense, with children drilling for test-savvy procedural ways to anticipate and answer questions, and, more important, that relying exclusively on multiple-choice tests fragments knowledge into small particles, consistent with behaviorism, long criticized as leading to the learning of trivial factoids.⁸⁴

This debate surfaced one of the most important anomalies in American testing. Its central tool, the standardized, multiple-choice question, reflects a long-criticized behaviorist psychology that is not consistent with our current education goals for all children, critical reading, reasoning, and thinking. According to Howard Gardner, by the

mid-1950's the behaviorist framework "had begun to come apart." Writing in the late 1980s, he concluded that "today the theoretical claims of behaviorism . . . are largely of historical interest." Still, the cognitive "revolution" that Gardner represented was not yet well enough known and understood to transform thinking about teaching, learning, and assessing. It had not penetrated mainstream concerns about test scores or to deflect the impulse to use high-stakes multiple-choice tests to drive education reform.⁸⁵

Nonetheless, by the end of the 1980s, cognitive science, with its more complex, process-oriented view of learning and thinking, was established as a discipline. University departments were re-formed or re-named, journals established, and researchers had developed a robust body of work. On the curriculum development front, the National Council of Teachers of Mathematics was sponsoring curriculum development that reflected cognitive science concepts, namely "constructivist" models of learning. Some states were beginning to experiment with alternative assessments like portfolios. These were the harbingers of important developments in the early 1990s.⁸⁶

The 1990's

The early 1990s witnessed the development of two fresh, influential ideas in school reform. One, standards-based reform, attracted a wide, centrist consensus and has been the mainstream reform strategy since the 1990s. The other, performance assessments, faced a number of daunting challenges. It has not become mainstream practice but is a crucial factor in thinking about the relationship of American test history to the reform of assessments as we move forward.

Standards-based reform (SBA) proposed that education systems (states, districts, nations) should develop content standards and performance standards. Content standards are learning objectives about what students should know and be able to do, specific to different school subjects. These standards should not dictate the details of curriculum but guide the creation of curriculum. Performance standards define what level and sort of competence will be required for a given content standard at different grade levels, and what scores will be labeled proficient, basic, or below basic. These standards would then be aligned with assessments, teacher preparation programs, and professional development.⁸⁷

Some advocates added opportunity-to-learn standards (OTL), to guarantee that a student would not be penalized for getting a low rating on subjects upon which she had not been taught. Advocates struggled with how to measure opportunity to learn; many conservatives opposed OTL standards on the grounds that they would hopelessly delay the implementation of the other standards. Others considered them a worthy idea but infeasible in terms of measurement. OTL receded into voluntary and then inactive status.⁸⁸

Standards-based reform enjoyed wide, bipartisan support. It developed a centrist consensus among reformers, state officials, Congress, and the Department of Education. Its appeal lay in its promise of coherence; the reform would be “systemic,” its parts “aligned.” Also, it promised to improve the goals of both excellence and equity. It would raise average achievement across the board, and it would reduce achievement gaps between rich and poor, as well as between race/ethnic groups.

The standards-based reform movement coincided with the advocacy of performance assessments, a movement widely promoted by assessment experts and school policy reformers. They aimed to develop assessments that more closely resembled real-world demonstrations of skills and knowledge. Such an assessment concept was very compatible with standards-based reform; indeed, if reformers hoped to realize the goal of high standards and a thinking curriculum, the dominance of standardized, multiple-choice tests would have to be reversed. Many of the people in the performance assessment movement of the early 1990s were reacting against the contagion of teaching to the test that resulted from placing high stakes on low-level, multiple-choice tests.⁸⁹

Performance assessment contrasted with the early twentieth-century legacy of multiple-choice achievement testing in several ways. It was more costly. Because more complex thinking was to be assessed, it was more complicated to understand. It was new and therefore threatened a long-established framework of standardization, scaling, norms, reliability, validity, and other technical features of the existing test framework. Also, it proposed to integrate assessments into the actual processes of teaching and learning. It would redress the balance in the purposes of assessment, with less focus on “summative” judgments for high stakes like grouping, placement and admissions, and more on “formative” assessment, with an emphasis on interacting with the student in the

instructional setting. That prospect would require much training and re-training of teachers, and it would leave in doubt whether the performance assessments could be used reliably to compare students and make other decisions about placement, credentialing, and school accountability.

Nonetheless, the advocates of performance assessment did not initially see these hurdles as insurmountable. The New Standards Project (NSP), under the direction of Professor Lauren Resnick of the University of Pittsburgh and Mark Tucker of the National Center on Education and the Economy at the University of Rochester, attempted a revision of standards-based-reform that emphasized more diverse modes of assessment, especially featuring performance assessment modes. Many in the standards movement thought this approach would help schools to assess students' higher-order skills, skills needed for the twenty-first century workplace, better than the focus on multiple-choice tests. NSP attracted 17 states and 10 large districts to try out the system. It was thus a project of national scope, and it generated a large amount of teacher professional development. Teachers were involved in creating the assessments as well as implementing them.⁹⁰

Performance assessment embodied principles shared by other assessment reformers in the 1990s. Appearing under the name "authentic assessment," "portfolio assessment," or "learner-centered assessment," many advocates in the U.S. and elsewhere advocated for assessments that were based more on real-life tasks, integrated into students' learning activities, and geared toward higher-level conceptual skills.⁹¹ Other reformers called for assessments that resonated better with a student's frame of reference, including diverse interests and backgrounds. This was yet another challenge to the national scope of standardized, uniform tests.⁹² The complexity of the new assessments made them quite different from the assessments to which teachers and principals were accustomed. And the substantial discussions about performance assessment heightened educators' consciousness that tests designed for particular purposes are often not well suited to other purposes, an abiding and difficult assessment issue.

Some conservative groups criticized the New Standards Project. More important, some education researchers emphasized the poor reliability of evaluations of students' work in such modes as portfolios, exhibitions and extended written work. Furthermore,

some said, the New Standards assessment process did not have as much transformative effect on teaching and curriculum as had been hoped. Although the Congressionally mandated National Council on Education Standards and Testing endorsed many of the principles embodied in the New Standards Project, they gained marginal traction in practice. The NSP waned.⁹³

Vermont provides an illustration of the obstacles. Vermont is a small state; as of 1990 its districts had the predominant role in education governance. There was no state assessment program. A new state superintendent of public instruction, Richard Mills, became interested in performance assessment and the New Standards Project. Traveling the state from border to border, he skillfully persuaded the districts to consider a move toward state testing, adopting the New Standards Reference Exam state-wide and implementing a system of portfolios. According to teachers, this proved to be well-liked and useful for diagnostic purposes; however, ratings by independent judges of the same piece of work were not very consistent, so the assessment results were not very “portable.” Pressure for more traditional tests came when Vermont’s Supreme Court mandated an equalization process across districts and demanded traditional evidence of achievement to compare districts. We note again that there is an equity side to the support of standardized normed tests to identify underperforming districts ill-serving poor or minority children.⁹⁴

Rick Mills departed Vermont to become New York State’s chief state school officer, and there he repudiated portfolios as having been a failed experiment. Mills’s predecessor had granted 28 districts in New York State waivers to substitute portfolios for standardized tests. When Mills discontinued these waivers, the *Daily Gazette* of Schenectady enthusiastically supported Mills’ decision, arguing that such innovations were an attack on factual knowledge and objective standards.⁹⁵

The powerful factors of economy, tradition, and accountability kept performance assessment and the testing of higher-order cognitive processes at the margins of the standards-based reform movement. Various groups tried to model and advocate high quality standards and assessments, notably ACHIEVE, a non-profit organization that was founded in 1996 and included among its prominent members governors, corporate executives and education policy makers. Nonetheless, as Lorrie Shepard wrote, although

“it may seem odd that the excellence movement, with its aversion for low standards, did not provoke a more thorough reexamination of the kinds of tests used to lead as well as measure the reform. . . even the new tests adopted in the mid 1980s were predominantly multiple-choice basic skills tests.”⁹⁶

Meanwhile, when President Bill Clinton entered the White House in January 1993, he was determined to re-establish a strong federal role in education. He had been active on behalf of education in Arkansas and was one of the leaders of discussions about education reform in the Education Commission of the States. Now he championed standards-based reform at the federal level. With Clinton’s Secretary of Education, Richard Riley, and Riley’s deputy, Marshall Smith, one of the chief theorists of standards-based education reform, the administration devised a bill called “Goals 2000,” which aimed to require states to develop standards and submit them for approval by a national council. The aim was to monitor the proposals for high quality but to encourage the states to develop different approaches and goals. The federal oversight features of Goals 2000 came under attack when a rejuvenated Republican Party came to power in the Congress after the elections of 1994. Before the Goals 2000 plan could be put into motion, Congress in 1995 abolished the national council that was designed to approve state standards.

The Clinton administration continued to press for standards in an advisory, partnership mode, while they used the reauthorization of the Elementary and Secondary Education Act to require states to evaluate Title I schools more rigorously, an initiative that met very slow compliance by the states. In its second term, Clinton education officials floated a proposal to have a Voluntary National Test and were again stopped by Republican opposition. There was considerable support to develop standards-based programs and continue education reform programs in many states, but little support for federal oversight. With regard to the ambitious education goals of Goals 2000, to raise average achievement and to narrow achievement gaps across class and racial lines, David Cohen and Susan Moffitt have provided a compelling argument that the political will and the resources were insufficient at all levels, federal, state and local, to have substantial success with the Title I mission. With regard to assessment, the pressure on districts to do more testing, generated by both the federal and state education agencies, certainly

succeeded in increasing the amount of testing. Whether the tests reflected more learning differed across states and was a matter of continual debate.⁹⁷

Into the Twenty-First Century

Like Bill Clinton, George W. Bush came to the White House determined to guide standards-based reforms from the federal level. This departed from his conservative base, but his involvement in education reform in Texas gave him determination. Dissidents in his party were mollified or pushed back, partly because they knew and appreciated the fact that he would implement conservative policies in other aspects of federal involvement. With the tragedy of September 11 as backdrop, the Bush administration skillfully got through the Congress the No Child Left Behind Act. They did so partly by negotiating the collaboration of liberal Senator Edward Kennedy, who saw the co-sponsorship as a way to insure more adequate funding for Title I and to implement the reporting of achievement scores disaggregated by class and race-ethnic categories of students. The law established an accountability system that had more negative sanctions than those of the Clinton administration. Definitions of satisfactory yearly progress were required; sanctions ultimately threatened “reconstitution” of a school, which could involve dismissal of principals or teachers, and, in any case, set the goals so implausibly high, that as many as half the school districts in the country would eventually be deemed failing. This caused much criticism, yet the NCLB regime survived both terms of the Bush administration, partly because of the strong centrist bipartisan history of the standards-based reform movement, which includes persistent support from civil rights advocates.

The civil rights aspect of SBR focused on the promise of more resources to underperforming schools and the disaggregation of achievement scores by race and class, in order to pinpoint pockets of underachievement across and within schools. Thus the provision in No Child Left Behind for frequent administration of achievement tests, the disaggregation of scores, as well as the negative consequences that flowed to districts that did not improve, were vigorously supported not only by Rod Paige, President Bush’s first Secretary of Education, but also by such civil rights stalwarts as William Taylor (executive director of the Citizens Commission on Civil Rights) and Christopher Edley

(director of the Harvard Civil Rights Project, now dean, University of California-Berkeley Law School).

The Obama administration declared the NCLB sanctions unreasonable and unproductive. The Department of Education gained a windfall influence over states and districts because of the financial collapse of 2008. The government's rescue operation provided a large stimulus package to the Education Department, which allowed the new Secretary to construct a "Race to the Top" that provided extra federal funds to states that complied with administration priorities about creating more charter schools and having teacher pay determined partly by student achievement test scores. In addition, states had to produce elaborate new systems of professional development, plans to save failing schools, and commitments other reform priorities. Lacking the political capital to mount a successful new education bill in the first term, the Department worked to minimize the consequences of NCLB sanctions through waivers to the states, in return for promises of reforms that would improve the test scores of disadvantaged students.

Despite the differences between the George W. Bush and the Barack Obama approach to the federal role in education, some common elements persist: both assume a framework of standards-based reform, and both see the federal government in a strong position of oversight. It will take time before a judgment can be made about the success of the Obama version of this reform. Myriad problems exist. Does the country have the political will to substantially reduce gaps in achievement levels? Do the players at each level have the capacity—the know-how and resources—to do the job? Can schools eradicate disadvantage without more broad social and economic reforms? In the meantime, is the emphasis on scores in basic reading and arithmetic narrowing and "dumbing down" the curriculum of the schools?

Reflections

One of the lessons from the past is that the now-traditional framework for standardized tests has tremendous staying power due to its economy, familiarity, and its seeming relevance to accountability of teachers and systems, plus the pervasive uses of standardized tests for the admission, classification, placement, and certification of students. Nonetheless, the performance assessment movement was very widespread and

raised very serious questions about the current emphasis in assessment on standardization and accountability and about the tension between this century-old framework and what scholars and classroom educators know about how children learn and develop.

Furthermore, the tremendous federal emphasis placed on tests in reading and math in the No Child Left Behind system has generated widespread frustration and opposition to excessive testing, test preparation, the narrowing of the curriculum through the focus on reading and math, and the fragmenting of knowledge represented by multiple-choice tests.

These qualms are not limited to academic experts but are widespread among teachers. In a passionate and moving opinion piece in the *New York Times*, Claire Needell Hollander recounted the emotional and perceptive reactions of students from disadvantaged circumstances in her reading enrichment classes in a New York City middle school. A reading specialist, Hollander designed these classes to give her kids the familiarity with literature that might help them survive and thrive in the more academic high schools of New York. But recently she has experienced ever-increasing pressure to abandon such literary offerings as *Macbeth*, *Of Mice and Men*, *Raisin in the Sun*, and *Catcher in the Rye* to drill students for the New York reading exams. Researching those tests, she found that among 600 multiple-choice items only one referred to a literary work of any kind.

She cannot prove whether her enrichment classes improve the students' reading or not. Some improved, others did not. But the reading tests are squeezing literature out of the curriculum, says Hollander, and they only ask questions about sentences that are complex but are empty of emotion and devoid of social problems. Thus, "we are sorting lower-achieving students into classes that provide less cultural capital than their already more successful peers receive in their more literary classes and depriving students who viscerally understand the violence and despair in Steinbeck's novels of the opportunity to read them." She pleads for the termination of standardized, culturally sanitized tests that are smothering the tradition of teaching literature in the schools. She recommends more extended written work, and evaluations which, however more subjective, would make the assessment more meaningful for teachers and students alike.⁹⁸

The Common Core:

Among the conditions required for states competing in the Race to the Top competition for federal funds was that participating states had to join multi-state consortia to develop assessments. Various small and large consortia popped up, but the dominant group was supported by the National Governors' Association and the Council of Chief State School Officers. They proposed a set of standards and matching assessments called the Common Core. Participation in these consortia raised obvious questions about the traditional control of curriculum by the individual states and districts. The usual fear about federal control—that the federal government would be dictating the content of schooling to local districts—was alleviated to some degree by the fact that the Common Core was proposed and developed as a set of *national*, not *federal*, standards and assessments, although the process was obviously initiated through federal pressure. Still, the governors and chief state school officers chose the people who are developing the standards and assessments, and many educators from the states are participating in those developments. Furthermore, any state can opt out if they are willing to forego the federal funds. (At the time of this writing, five states had not joined.)

In many states, people voiced the traditional opposition to incursion on state and local control, but a growing number of administrators, researchers, and policy makers had come to see the economy of scale in the process of states banding together to develop complicated standards and assessments. The convictions of such educators, plus the lure of federal funds in a time of stressed budgets, carried the day. Their faith is that the states collectively have the capacity to create and implement first-class standards and assessments, while the states individually have uneven capacity to do so. As a result, the participating states are implicitly agreeing to modify curriculum and teacher training to conform to the Common Core program.⁹⁹

The leaders of Common Core aspire to first-rate education and first-rate standards. Yet many of the assessments best suited to higher-order thinking and reasoning processes either require much more testing time or computer assisted administration, both of which have aroused practical complaints in the districts and states. The Common Core assessment programs were still in development as this essay was written. It will take several years before we can judge the usefulness and success of the

Common Core in general and the two main assessment systems from which states are choosing. The current debates reflect persistent tensions between formative and summative assessment, between getting information that can help improve teaching and learning, which takes more testing time and more difficult rating processes, and getting information about how “much” students have learned, which risks diminishing the assessment of higher-end capacities needed for a twenty-first-century education¹⁰⁰

Lessons from history?

History cannot predict the future, nor tell us with any certainty what policy options to choose. But it can trace the origins of important ideas and explore the reasons why some get deeply embedded in educational practice and why some do not.¹⁰¹ Here, then, is a recapitulation of the larger historical generalizations one can draw from the history presented in this essay.

First, this is not a simple story of right and wrong, of bad folks and good folks. It is a complicated struggle between competing values. This essay has emphasized the costs involved in having such a central focus on standardized multiple-choice tests of basic skills. But there are important equity goals and accountability goals that are served by pinpointing who is doing well, among individual students or across groups of students, or in comparing teachers or schools. The solution must lie in some new balancing of modes and uses of assessments.

Second, there is a great disjunction between our stated goals of higher-order thinking, learning how to learn, student-initiated learning, and cultural diversity on the one hand, and the focus of testing today on basic skills and standardized, multiple-choice assessments. The aspiration to teach *all* students high-level literacy and numeracy and the ability to think critically was a new, unprecedented idea in the history of education, originating after World War II in the United States and elsewhere. It is a grand, challenging hope. One significant obstacle is learning driven by standardized, multiple-choice tests.¹⁰²

Third, the present testing practices have powerful support because standardized test scores have become a metric for what a good education is, and thus their usefulness in placement, accountability and program evaluation. They are powerful because

standardized, multiple-choice tests are less costly and more efficient compared to the more complex, more subjective and higher-level assessments. Furthermore, there is widespread popular belief in the priority of factual knowledge as the proper and primary goal of schools, a belief that is shared by many people who themselves have had very elegant liberal educational experiences.

There are a number of other important developments in the history of assessments, not all of them detailed in this paper. The uses of tests have changed over time. Nineteenth-century tests related primarily to individual students' abilities and the consequences attached to those judgments, with a minor role in rating teachers or schools' effectiveness. Gradually these functions melded with, and to some degree were overshadowed by, ratings of schools' performance, states' performance, and the effectiveness of programs. Many technical and technological changes have occurred in the conception, construction and uses of tests: from machine scoring, to item response theory, to elaborate discussions about the concept of validity, and, more recently, rapidly multiplying possibilities for computer assisted assessment. The basic concept of intelligence has undergone great change in mainstream discussions: from a widespread hereditarian view that saw IQ as a permanent, unitary feature of an individual's mental capacity, to a mainstream view that, while heredity is still a contributing component of initial intelligence, mental abilities are malleable and multiple. When Richard Herrnstein and Charles Murray's book, *The Bell Curve* (1994), revived visions of high IQ whites and Asians barricaded against increasingly low IQ low-income, largely minority people, and urged policies discouraging child-bearing among the poor, it caused another very public debate but could not reverse the change in public opinion that had occurred over the decades.¹⁰³ Amidst all of these changes, one continuity stands out: the dominance of multiple-choice, standardized tests has not changed in a century of testing practice.

The current moment in assessment reform

Current assessment reforms promise new and creative ways to integrate assessment with teaching and learning, update assessment practices to comport with what we now know about learning and about the potential of digital tools for learning and assessment. These could contribute to a richer, more active, more thoughtful, more

individualized, and more cross-cultural education for our children. But because these assessment reforms are also more costly, because they are more difficult for lay people to understand, and because they challenge not only the current relationship between assessment and learning but also the dominance of accountability, the advocates of such reforms will have to develop concrete descriptions of what is possible, why it is important, and why it is urgent that such changes be made.

Those descriptions must be accessible to educators who are not experts in assessment. They should make a clear argument about how a new system of assessments can address the currently dominant goals of accountability of staff and systems as well as the processes of admission, classification, placement, and certification of students. On the other hand, the reform must be strong enough and sufficiently thorough to displace the negative, dysfunctional features of current testing practice: the pervasive attention to accountability, the fragmentation of knowledge, the focus on low-level skills, the narrowing of curriculum, the stifling of creative teaching, the indifference to the cultural and life-experience frames of individual students, and the diminution of active participation of students in their education.

Hopeful signs:

As a historian and a consultant to the Gordon Commission on the Future of Educational Assessment, I am certainly not the person to make such an argument. The distinguished chair, co-chair, and members of the commission will make it. But having observed their progress to date, I am convinced that we are at an important moment, hopefully an auspicious moment, in the history of assessment. I believe that a persuasive argument can be made for reform. I say that despite the evidence presented in this essay about the durability and pervasiveness of current testing practices and the difficulties encountered by proposals for performance assessments over the past twenty-five years.

During the first dozen years of the 21st century, many strong contributions to support such a reform have been made. As the century began, the National Research Council's Committee on Developments in the Science of Learning (1999) surveyed the key contributions of the cognitive revolution in learning theory, of brain science, and of research on effective teaching in the disciplines. The National Research Council's

Committee on the Foundations of Assessment (2001) surveyed what was at stake and urged a shift to multiple sources of assessment, more local and less external assessment, strengthening the cognitive coherence across curriculum, instruction and assessment, and many other elements of a reform direction. Scholars like Michael Martinez have helped to redefine intelligence in a humane and creative direction.¹⁰⁴

Many other signal contributions could be mentioned. Many advances have been made in computer assisted instruction and assessments involving higher level skills, collaborative work, and other rich activities. Excellent and detailed papers have been written for the Gordon Commission. Finally, there now exists a detailed model of how a comprehensive argument can be made for a specific program of assessment reform. In “An American Examination System,” Lauren Resnick and Larry Berger address the problem, their solution, its relationship to the Common Core Standards, the uses of assessments in their system for both formative and summative purposes, cost factors, the uses of technology, and other issues. Whether all the specifics are just right, or will gain sufficient consensus, is not the point. The point is that the Resnick-Berger prospectus provides a model for how the argument can be constructed.¹⁰⁵

Assessment practices we have inherited, and which are deeply embedded in our education system, are out of date, and they are out of step with our best educational aspirations. We can use this history to apply our *knowledge*, not just our traditions, to the future.

Endnotes

¹ The work of Daniel P. Resnick is foundational for the policy context of American testing history. See especially, Resnick, “*Minimum Competency Testing Historically Considered*,” *Review of Research in Education* 8 (1980): 3—29; Resnick, “History of Educational Testing,” in *Ability Testing: Uses, Consequences, and Controversies. Part II: Documentation Section* ed. Alexandra K. Wigdor and Wendell R. Garner (Washington, National Academy Press, Committee on Ability Testing, National Research Council, 1982): 173—191; and Daniel P. Resnick and Lauren B. Resnick, “Performance Assessment and the Multiple Functions of Educational Measurement,” in *Implementing Performance Assessment: Promises, Problems and Challenges*, Michael B. Kane and Ruth Mitchell eds. (Mahwah, New Jersey, Laurence Erlbaum, 1996): 23—38. See also,

Raymond E. Fancher, *The Intelligence Men: Makers of the IQ Controversy* (New York: W. W. Norton: 1985); *Psychological Testing and American Society, 1890—1930*, ed. Michael M. Sokal (New Brunswick, N. J.: Rutgers University Press, 1987); U. S. Congress, Office of Technology Assessment, *Testing in American Schools: Asking the Right Questions*, Michael J. Feuer, Project Director, OTA-SET-519 (Washington, DC: U. S. Government Printing Office, February, 1992): 3—9, 81—99, 103—131; and Paul Davis Chapman, *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890—1930* (New York: New York University Press, 1988). Robert Rothman has written two valuable books on the recent decades: *Measuring Up: Standards, Assessment, and School Reform* (San Francisco: Jossey-Bass, 1995) and *Something in Common: The Common Core Standards and the Next Chapter in American Education* (Cambridge, Mass.: Harvard Education Press, 2011). For the recent decades also see Lorrie A. Shepard, “A Brief History of Accountability Testing, 1965—2007 in *The Future of Test-Based Accountability*, ed. Katherine E. Ryan and Lorrie A. Shepard (New York: Routledge, 2008): 25—46; and Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, Mass.: Harvard University Press, 2008), chapter 4.

² Paul Black, *Testing: Friend or Foe? Theory and Practice of Assessment and Testing* (London: Falmer Press, 1998): 7—11; Philip H. DuBois, *A History of Psychological Testing* (Boston: Allyn & Bacon, 1970): 3—5; David L. McArthur, “Educational Testing and Measurement: A Brief History,” Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles, CSE Report No. 216, 1983): 1.

³ Michael Feuer, personal communication, January, 2013.

⁴ On the relationship of classroom assessments to external, standardized assessments today, see Lorrie A. Shepard, “Classroom Assessment,” in *Educational Measurement Fourth Edition*, ed. Robert L. Brennan (Westport, Conn.: Praeger, 2006): 623—646.

⁵ William J. Reese, *Testing Wars: The Untold Story* (Cambridge, Mass.: Harvard University Press, forthcoming), Chapter 1.

⁶ Reese, *Testing Wars*, Chapter 2. For the ideology, politics, and educational program of the Whig Party, see Carl F. Kaestle, *Pillars of the Republic: Common Schools and American Society, 1780—1860* (New York: Hill & Wang, 1983).

⁷ Reese, *Testing Wars*, Chapters Three and Four

⁸ The quote is in Reese, *Testing Wars*, p. 208; for Wells, pp. 220—222, for test-prep materials, pp. 207—208, 262.

⁹ For Omaha and Portland, see Reese, *Testing Wars*, pp. 203—204; for Cincinnati, pp. 228—229; for the increase in writing, p. 235.

¹⁰ Reese, *Testing*, p. 248—249 for the critics; the quotation is from p. 257.

¹¹ Reese, *Testing Wars*, 283.

¹² On Pearson, see DuBois, *History of Psychological Testing*, 14—15; on Spearman, see Fancher, 84—98.

¹³ See Lauren B. Resnick and Daniel P. Resnick, “Assessing the Thinking Curriculum: New Tools for Educational Reform,” in B. R. Gifford & M. C. O’Connor, eds., *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction* (Boston: Kluwer, 1992): 37—75.

- ¹⁴ Fancher, *Intelligence Men*, 41—49; Michael M. Sokal, “James McKeen Cattell and Mental Anthropometry: Nineteenth-Century Science and Reform and the Origins of Psychological Testing,” in Sokal, *Psychological Testing*, 21—40.
- ¹⁵ Fancher, *Intelligence Men*, 49—69.
- ¹⁶ Fancher, *Intelligence Men*, 70—83.
- ¹⁷ Henry Herbert Goddard, *The Kallikak Family: A Study in the Heredity of Feeble-Mindedness* (New York: Macmillan, 1912).
- ¹⁸ For Tyack’s depiction of the administrative progressives, see David B. Tyack, *The One Best System: A History of American Urban Education* (Cambridge, MA.: Harvard University Press, 1974).
- ¹⁹ Paul Davis Chapman, *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890—1930* (New York: New York University Press, 1988): 22—24, 28, 32.
- ²⁰ Daniel J. Kevles, “Testing the Army’s Intelligence: Psychologists and the Military in World War I,” *Journal of American History* 55:3 (December, 1968): 565—581.
- ²¹ J. C. Bell, “Recent Literature on the Binet Tests,” *Journal of Educational Psychology* 3 (1912): 101—110, cited in Gerard Giordano, *How Testing Came to Dominate American Schools: A History of Educational Assessment* (New York: Peter Lang, 2005): 18.
- ²² For these and other trends in migration by region, see Philip Martin and Elizabeth Midgley, “Immigration: Shaping and Re-Shaping America,” *Population Bulletin* 58:2 (2003): 13.
- ²³ Robert H. Wiebe, *The Search for Order, 1877—1920* (New York: Hill and Wang, 1967); Ellis W. Hawley, *The Great War and the Search for a Modern Order: A History of the American People and Their Institutions, 1917—1933* (New York: St. Martin’s Press, 1979); Alan Trachtenberg, *The Incorporation of America: Culture and Society in the Gilded Age* (New York: Hill and Wang, 1982); James R. Beniger, *The Control Revolution: Technological and Economic Origins of the Information Society* (Cambridge, Mass.: Harvard University Press, 1986); Alfred D. Chandler Jr., *The Visible Hand: The Managerial Revolution in American Business* (Cambridge, Mass: Harvard University Press, 1977; and Samuel Haber, *Efficiency and Uplift: Scientific Management in the Progressive Era, 1890—1920* (Chicago: University of Chicago Press, 1964). For the effects of corporate capitalism on print culture and higher education, see Carl F. Kaestle and Janice A. Radway, “A Framework for the History of Publishing and Reading in the United States, 1880—1940,” and Janice A. Radway, “Learned and Literary Print Cultures in an Age of Professionalization and Diversification,” both in *Print in Motion: The Expansion of Publishing and Reading in the United States, 1880 to 1940* ed. Carl F. Kaestle and Janice A. Radway (Chapel Hill: University of North Carolina Press, 2008), chapters 1 and 11.
- ²⁴ See John Rury, *Education and Women’s Work: Female Schooling and the Division of Labor in Urban America, 1870—1930* (Albany, N.Y.: State University of New York Press, 1991); and Raymond E. Callahan, *Education and the Cult of Efficiency: A Study of the Social Forces that have Shaped the Administration of the Public Schools* (Chicago: University of Chicago Press, 1962).
- ²⁵ G. M. Whipple, “The National Intelligence Tests,” *Journal of Educational Research* 4 (1921): 16

- ²⁶ Office of Technology Assessment, *Testing in American Schools*, 122.
- ²⁷ Walter Lippmann, “The Reliability of Intelligence Tests,” *New Republic* (November 8, 1922): 276; Lippmann, “The Abuse of the Tests,” *New Republic* (November 15, 1922): 297.
- ²⁸ Walter Lippmann, *Drift and Mastery: An Attempt to Diagnose the Current Unrest* (New York: Mitchell Kennerley, 1914; Spectrum paperback edition, Englewood Cliffs, N.J. : Prentice-Hall, 1961): 151
- ²⁹ *Plessy v. Ferguson* 163 U.S. 537 (May 18, 1896), on line at Lexis-Nexus Legal, accessed on July 17, 2012; *Encyclopedia Britannica* Eleventh Edition (1911): volume 19, p. 344, cited in David Brion Davis, “Should You Have Been an Abolitionist?” *New York Review of Books* 59:11 (June 21, 2012): 58; Ellwood Cubberley, *Public Education in the United States: A Study of American Educational History* (Boston: Houghton Mifflin, revised edition, 1934): 700.
- ³⁰ See Richard E. Nisbett, et alia, “Intelligence: New Findings and Theoretical Developments,” *American Psychologist* 67:2 (February—March, 2012): 130—159; and Eka Roivainen, “Gender Differences in Processing Speed: A Review of Recent Research” *Learning and Individual Differences* 21 (2011): 145—149. An accessible piece on sex differences in cognitive abilities is Sharon Begley, “Gray Matters,” *Newsweek* (March 27, 1995), subsequently published in *The Biological Basis of Human Behavior: A Critical Review* ed. R. W. Sussman (Upper Saddle River, NJ, Prentice Hall, 2nd edition, 1999): 383—338, still basically up-to-date.
- ³¹ James McKeen Cattell, “The School and the Family,” *Popular Science Monthly* 74 (January 1909): 91, 92 cited in Rosalind Rosenberg, *Beyond Separate Spheres: Intellectual Roots of Modern Feminism* (New Haven, Conn.: Yale University Press, 1982): 89.
- ³² Edward Thorndike, quoted in Clarence J. Karier, “Testing for Order and Control in the Corporate Liberal State,” *Educational Theory* 22:2 (April, 1972): [page # pending]
- ³³ See Helen Thompson Woolley, *The Mental Traits of Sex* (Chicago: University of Chicago Press, 1903; Leta Stetter Hollingworth, “Sex Differences in Mental Traits,” *Psychological Bulletin* 13 (October, 1916): 377—385, and the discussion of this work in Rosenberg, *Beyond Separate Spheres*, 57, 68—89, 95—97, 103, 107.
- ³⁴ Richard Hofstadter, *Social Darwinism in American Thought* (Boston: Beacon Press, 1944, revised Beacon Press paperback, 1955): 46, 178, 180, 202—204; on the decline of the eugenics movement, see Daniel Kevles, *In the Name of Eugenics: Genetics and the Uses of Human Heredity* (New York: Alfred Knopf, 1975): 128—147, 164—166; and Mark H. Haller, *Eugenics: Hereditarian Attitudes in American Thought* (New Brunswick, NJ: Rutgers University Press, 1963): 179—183.
- ³⁵ For Snedden, see Walter H. Drost, *David Snedden and Education for Social Efficiency* (Madison, Wis.: University of Wisconsin Press, 1967): 168—169.
- ³⁶ Franklin Bobbitt, “Some General Principles of Management Applied to the Problems of City-school Systems,” in *The Supervision of City Schools* (Bloomington, Ill.: Twelfth Yearbook of the National Society for the Study of Education, Part I, Public School Publishing Company, 1913): 7—96.
- ³⁷ Edward L. Thorndike, *Educational Psychology: Volume II, The Psychology of Learning* (New York: Teachers College, Columbia University, 1913): 4. See Geraldine

M. Joncich, ed., *Psychology and the Science of Education: Selected Writings of Edward L. Thorndike* (New York: Teachers College Press, 1962) and Joncich, *The Sane Positivist: A Biography of Edward L. Thorndike* (Middletown, Conn.: Wesleyan University Press, 1968).

³⁸ David Tyack and Elisabeth Hansot, *Managers of Virtue: Public School Leadership in America, 1820—1980* (New York: Basic Books, 1982):129—140.

³⁹ The test examples are cited in William J. Reese, *The Origins of the American High School* (New Haven, CT: Yale University Press, 1995): 149, 151.

⁴⁰ Walter S. Monroe, Charles W. Odell, M. E. Herriott, Max D. Engelhart, and Mabel R. Hull, *Ten Years of Educational Research, 1918—1927* (Urbana, Ill., University of Illinois, College of Education, Bureau of Research, Bulletin No. 42, 1928): Chapter IV, “Research in Educational Measurement, Part 1. “Before 1918,” 90—93; Leonard P. Ayres, “A Measuring Scale for Ability in Spelling,” (New York: Russell Sage Foundation, 1915): 12—20.

⁴¹ On Otis and multiple choice, see DuBois, *History of Psychological Testing*, 59, 73.

⁴² Gerard Giordano, *How Testing Came to Dominate American Schools: The History of Educational Assessment* (New York: Peter Lang, 2005): 93—94, and more generally, chapters 2 and 4.

⁴³ See George D. Strayer, “Report of the Committee on Tests and Standards of Efficiency in Schools and School Systems: A Brief Statement Concerning the Purpose, Nature, and Conduct of School Surveys,” *Journal of the Proceedings and Addresses of the 52nd Annual Meeting of the National Education Association* (Ann Arbor, Mich: National Education Association, 1914): 302—310; see also *The Measurement of Educational Products* ed. Guy Montrose Whipple (Bloomington, Ill: Seventeenth Yearbook of the National Society for the Study of Education, Part II, Public School Publishing Company, 1918) on the status and details of textbook production in 1918.

⁴⁴ Fred Newton Scott, “Efficiency for Efficiency’s Sake,” *The School Review* 23:1 (January, 1915): 36; Leonard P. Ayres, “Measuring Educational Processes through Educational Results,” *School Review* 20 (1912): 302, 307. Tracy Steffes, *School, Society, and State: A New Education to Govern Modern America, 1890—1940* (Chicago, Ill.: University of Chicago Press, 2012): 31—44.

⁴⁵ Carl C. Brigham, *A Study of American Intelligence* (Princeton, N. J.: Princeton University Press, 1923); see Nicholas Lemann, *The Secret History of the American Meritocracy* (New York: Farrar, Straus and Giroux, 1999): 29—32.

⁴⁶ Lemann, *Big Test*, 38—66; See also Claude M. Fuess, *The College Board: Its First Fifty Years* (New York: Columbia University Press, 1950): 100—119, 182—193.

⁴⁷ On the criticisms and on Educational Testing Service’s responses to them, see Lemann, *The Big Test*, 221—232, 268—277. For more recent decisions by colleges to de-emphasize the importance of SAT scores in admissions, see Ronald J. Coleman, “Stratification, Inequality, and the SAT: Toward an SAT-Optional Movement,” *Georgetown Journal on Poverty Law & Policy* 18:3 (2011) 507—531.

⁴⁸ *Digest of Education Statistics* (Washington, D.C.: Center for Education Statistics, 1970 edition). See also Martin Trow, “The Second Transformation of American Secondary Education,” *International Journal of Comparative Sociology* 2:1 (1966): 144—166.

⁴⁹ Lee J. Cronbach, “Five Decades of Public Controversy Over Mental Testing,” *American Psychologist* (January, 1975): 7—8. See also Carl F. Kaestle, “Public Schools and the Public Mood” *American Heritage* (February 1990): 66-81.

⁵⁰ On the NDEA see Wayne J. Urban, *More Than Science and Sputnik: The National Defense Education Act of 1958* (Tuscaloosa: The University of Alabama Press, 2010). On the early twentieth-century origins of NDEA’s ideology, see David Gamson, “From Progressivism to Federalism: The Pursuit of Equal Educational Opportunity, 1915—1965,” in *To Educate a Nation: Federal and National Strategies of School Reform* ed. Carl F. Kaestle and Alyssa E. Lodewick (Lawrence: University Press of Kansas, 177—201). On the inaccuracy of the admiring view of Soviet Schools, see Susan Jacoby, *Inside Soviet Schools* (New York: Hill & Wang, 1974).

⁵¹ James B. Conant, *The American High School Today* (New York: McGraw-Hill Book Company, 1959).

⁵² *Griggs v. Duke Power Company* 401 US 424 (1971)

⁵³ *Singleton v. Jackson Municipal Separate School System* 419 F 2d 121 (1969); *Larry P. v. Riles* 343 F. Supp 1036 (1972). The *Griggs*, *Singleton*, and *Larry P.* cases are discussed in *Ability Testing: Uses, Consequences and Controversies. Part I. Report of the Committee* ed. Alexandra K. Wigdor and Wendell R. Garner (Washington, D.C.: Committee on Ability Testing, National Research Council, National Academy Press, 1982): 97—110. Further cases are discussed in Patricia Hollander, “Legal Context of Educational Testing,” in *Ability Testing: Uses, Consequences, and Controversies. Part II: Documentation Section*, ed. Alexandra K. Wigdor and Wendell R. Garner (Washington, D.C.: Committee on Ability Testing, National Research Council, National Academy Press, 1982): 195—231. See also Alexandra K. Wigdor and John A. Hartigan, *Fairness in Employment Testing: Validity, Generalization, Minority Issues, and the General Aptitude Battery* (Washington, D.C.: National Research Council, National Academy Press, 1989).

⁵⁴ The debates concerning the test score gap, including the controversy over the role of heredity in academic achievement, is treated in *The Black-White Test Score Gap* ed. Christopher Jencks and Meredith Phillips (Washington, D.C.: Brookings Institution Press, 1998).

⁵⁵ Francis Keppel to Rice Clemow, Oct `7, 1963. Office Files of the Commissioner of Education, Box 94, Folder 6, NARA2.

⁵⁶ Francis Keppel, oral history interview, July 18, 1968, transcript, pp. 29—30, Box Ac66-1, LBJ Library.

⁵⁷ Lyle V. Jones, “A History of the National Assessment of Educational Progress and Some Questions About its Future,” *Educational Researcher* 25:7 (October, 1996): 15—22.

⁵⁸ Harold C. Hand, “National Assessment Viewed as The Camel’s Nose,” *Phi Delta Kappan* 47:1 (September, 1965): 8—13.

⁵⁹ Ralph W. Tyler, “Assessing the Progress of Education,” in *ibid.*, 13—16.

⁶⁰ Francis Keppel to Joseph Califano, August 10, 1963, White House Central Files, Education, Box 2, Folder 1, LBJ Library.

⁶¹ Keppel to Senator Roman Hruska, Sept 29, 1965, Box 229, Congressional A-Z, July—December, 1965, Office Files of the Commissioner of Education, NARA2; and Wilbur Cohen to Everett Dirksen, *ibid.*

⁶² See Carl F. Kaestle, *Uncertain Mandate: Formative Years of The Federal Role in Education* (in progress)

⁶³ “School Crackdown Starts on Local Schools,” *U. S. News and World Report*, October 18, 1965.

⁶⁴ Gene Currivan, “Keppel Recommends Overhaul of State Education Organizations and Policies,” *New York Times* February 15, 1966): 24.

⁶⁵ Jones, “History of National Assessment”; Lauren B. Resnick, “Reflections on the future of NAEP: Instrument for Monitoring or for Accountability?” Center for the Study of Evaluation, Technical Report # 499 (Los Angeles, CA: National Center for Research on Evaluation, February, 1999); National Academy of Education, *Assessment in Transition: Monitoring the Nation’s Educational Progress* Robert Glaser and Robert Linn, panel chairmen, Panel on the Evaluation of the NAEP Trial State Assessment (Stanford, Cal.: The National Academy of Education: 1997); National Research Council, Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment, National Research Council, *Grading the Nation’s Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress* ed. James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell (Washington, D.C :National Academy Press, 1999)

⁶⁶ On Robert Kennedy’s shabby treatment of Johnson at meetings of the Commission on Equal Employment Opportunity, see Dallek, *Flawed Giant*, 35—36.

⁶⁷ On the influence of the President’s Committee on Juvenile Delinquency on Robert Kennedy’s views of poverty, see Edward R. Schmitt, *President of the Other America: Robert Kennedy and the Politics of Poverty* (Amherst, Mass.: University of Massachusetts Press, 2010): 95—102. Joseph Alsop, “Matter of Fact: the Kennedy Style,” *Washington Post* (March 3, 1965): A17.

⁶⁸ Francis Keppel, “Poverty: Target for Education,” address to the Americana Association of School Administrators, Atlantic City (February 15, 1964): 5, 7, 9—11, 13, Office Files of the Commissioner of Education, Keppel Speeches, January—May, 1964, Box 168, NARA2.

⁶⁹ Jeff Shesol, *Mutual Contempt: Lyndon Johnson, Robert Kennedy, and the Feud That Defined a Decade* (New York: Norton, 1997): 239—240; Rowland Evans and Robert Novak, “Inside Report: Robert Kennedy and the School Bill,” *Washington Post* (February 23, 1965): A 17; Joseph Alsop, “Kennedy Style.”

⁷⁰ Schmidt, *President of the Other America*, 116—117. Eligibility for Title I in the enacted legislation requires that the state determine “that effective procedures, including provision for appropriate objective measurements of educational achievement, will be adopted for evaluating, at least annually, the effectiveness of the programs in meeting the special educational needs of educationally deprived children” and “that the local educational agency will make an annual report and such other reports to the State Educational agency, in such form and containing such information, as may be reasonably necessary to enable the State educational agency to perform its duties until this title, including information relating to the educational achievement of students participating in

programs carried out under this title.” *Elementary & Secondary Education Act* (P.L./ 89—10), *United States Statutes at Large* Volume 79, p. 31, on line at <http://www.necticlp.org/files/40646763.pdf>. Accessed February 28, 2012.

⁷¹ Oral history interview, Emerson Elliott, by the author, Washington, D.C., October 21, 2008; Laurence E. Lynn, Jr., “Reform of the Federal Government: Lessons for Change Agents,” manuscript, paper for the LBJ Centennial Symposium, LBJ School of Public Affairs, December 4—5, 2008, pp. 1—11, forthcoming in *LBJ’s Neglected Legacy* ed. Laurence Lynn, Norman J. Glickman and Robert H. Wilson (Austin, Tex.: University of Texas Press)

⁷² Lynn, “Reform of the Federal Government,” 9.

⁷³ Gerald Grant, “Shaping Social Policy: The Politics of the Coleman Report,” *Teachers College Record* 75::1 (September, 1973): 17—54.

⁷⁴ James Coleman, *et al. Equality of Educational Opportunity* (Washington, D.C.: Government Printing Office, 1966). An interesting summary and review is Christopher Jencks, “Education: The Racial Gap,” *New Republic* (October 1, 1966): 21—26.

⁷⁵ Grant, “Shaping Social Policy,” 22—29, 32. See also Daniel P. Moynihan, “Sources of Resistance to the Coleman Report,” *Harvard Educational Review* 38:1 (Winter, 1968): 23—36, and James S. Coleman, “Toward Open Schools,” *The Public Interest* (Summer, 1966): 20—27) for Coleman’s own startling policy suggestions.

⁷⁶ Two prominent examples of commentary are *On Equality of Educational Opportunity* ed. Frederick Mosteller and Daniel Patrick Moynihan (New York: Random House, 1972); and Christopher Jencks *et alia, Inequality: A Reassessment of the Effect of Family and Schooling in America* (New York: Basic Books, 1972). A recent example of research that re-analyzed the Coleman data and found greater impact of school resources on achievement, is Geoffrey D. Borman and N. Maritza Dowling, “Schools and Inequality: A Multilevel Analysis of Coleman’s Equality of Educational Opportunity Data,” *Teachers College Record* 112 (2010): 1201—1246.

⁷⁷ Resnick, “Minimum Competency Testing,” 17—23.

⁷⁸ Katherine A. McDermott, *High-Stakes Reform: The Politics of Educational Accountability* (Washington, D.C.: Georgetown University Press, 2011). See also Robert L. Linn, “Assessments and Accountability,” *Educational Researcher* 29:2 (March, 2000): 4—16; and Lorrie A. Shepard, “A Brief History of Accountability Testing, 1965—2007,” in *The Future of Test-Based Accountability* (New York: Routledge, 2008): 25—46

⁷⁹ Arthur Jensen, “How Much Can We Boost IQ and Academic Achievement,” *Harvard Educational Review* 39:1 (Spring, 1969): 1—123.

⁸⁰ Richard Herrnstein, “IQ,” *The Atlantic Monthly* (September, 1971). The academic response to Jensen began with a special issue of the *Harvard Educational Review* 39:2 (Summer, 1969), with articles by Jerome Kagan, J. McVicar Hunt, James F. Crowe, Carl Bereiter, David Ellkind, Lee J. Cronbach, and William F. Brazziel.

⁸¹ National Commission on Excellence in Education, *A Nation at Risk. The Imperative for Education Reform.* (Washington, D.C.: U.S. Government Printing Office, 1983); for commentary on the test score decline debate, see Carl F. Kaestle and Lawrence Stedman, “The Test Score Decline is Over: Now What?” *Phi Delta Kappan* 67 (November, 1985): 204-210.

⁸² On accountability, see Kathryn A. McDermott, *High-Stakes Reform: The Politics of Educational Accountability* (Washington, D. C.: Georgetown University Press, 2011). There is a growing and now robust literature on the history of federal policy since the 1983 *Nation At Risk* report; see Maris A. Vinovskis, *From A Nation at Risk to No Child Left Behind: National Education Goals and the Creation of Federal Education Policy* (New York: Teachers College Press, 2009); Diane Ravitch, *National Standards in American Education: A Citizen's Guide* (Washington, D.C.: The Brookings Institution, 1995); John F. Jennings, *Why National Standards and Tests? Politics and the Quest for Better Schools* (Thousand Oaks, Cal.: Sage Publications, 1998); Patrick J. McGuinn, *No Child Left Behind and the Transformation of Federal Education Policy, 1965—2005* (Lawrence: University Press of Kansas, 2006); Paul Manna, *School's In: Federalism and the National Education Agenda* (Washington, D.C.: Georgetown University Press, 2006); and *To Educate a Nation: Federal and National Strategies of School Reform*, ed. Carl F. Kaestle and Alyssa E. Lodewick (Lawrence: University Press of Kansas, 2007).

⁸³ Chris Pipho, "Stateline: Some Issues Won't Go Away," *Phi Delta Kappan* 68:1 (June, 1987): 726; on the reauthorization of the elementary and secondary bill, now called the Education Consolidation and Improvement Act, see Christopher T. Cross, *Political Education: National Policy Comes of Age* (New York: Teachers College Press, 2004):88—89.

⁸⁴ W. James Popham, "The Merits of Measurement-Driven Instruction," *Phi Delta Kappan* 68: 9 (May, 1987): 679—682; Lorrie A. Shepard, "Should Instruction Be Measurement-Driven? A Debate," paper given at the annual meeting of the American Educational Research Association, New Orleans, April, 1988 (text provided to me by the author).

⁸⁵ Howard Gardner, *The Mind's New Science: A History of the Cognitive Revolution* ((New York: Basic Books, paperback edition with new epilogue, 1987): 110.

⁸⁶ National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, Va.: NCTM, 1989); for the increasing criticism of standardized tests, and early hints of alternative assessments before 1990, see Rothman, *Measuring Up*, Chapters 3 and 4.

⁸⁷ See Marshall Smith and Jennifer O'Day, "Systemic School Reform" in *The Politics of Curriculum and Testing* ed. Susan H. Fuhrman and B. Malen (London: Taylor & Francis, 1990); Lauren B. Resnick and Daniel P. Resnick, "Assessing the Thinking Curriculum: New tools for Educational Reform," in *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction* ed. B. R. Gifford and M. C. O'Connor (Boston: Kluwer Academic Publishers, 1992): 37—75; Milbrey W. McLaughlin and Lorrie A. Shepard, *Improving Education through Standards-Based Reform. A Report of the National Academy of Education Panel on Standards-Based Reform* (Stanford, CA: National Academy of Education, 1995); and Margaret E. Goertz, Robert E. Floden, and Jennifer O'Day, *Studies of Education Reform: Systemic Reform, Volume I: Findings and Conclusions* (New Brunswick, N. J.: Consortium for Policy Research in Education, July, 1995).

⁸⁸ Andrew C. Porter, "School Delivery Standards," *Education Researcher* 22:5 (1993): 13—29; and Porter, "The Uses and Misuses of Opportunity to Learn Standards," *Education Researcher* 24:1 (January—February, 1995): 21—27.

⁸⁹ I witnessed the early enthusiasm for high standards and new types of assessments as a visitor to the Pew Forum on Elementary and Secondary Education (sponsored by the Pew Charitable Trust). On the critique of “teaching the test,” see Resnick and Resnick, “Assessing the Thinking Curriculum,” and Shepard, “Should instruction Be Measurement-Driven?” Also see Linda Darling-Hammond, “Performance-Based Assessment and Educational Equity,” *Harvard Educational Review* 64:1 (Spring, 1994): 5—30; Douglas A. Archbald and Fred M. Newman, *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School* (Reston, Va.: National Association of Secondary School Principals, 1988); Robert L. Linn, “Educational Assessment: Expanded Expectations and Challenges,” *Educational Evaluation and Policy Analysis* 15:1 (Spring, 1993): 1—16.

⁹⁰ Elizabeth Spaulding, “New Standards Project and the English Language Arts Portfolio: Report on Process and Progress,” *The Clearing House* 68:4 (March—April, 1995): 219—223; Interview, “On the New Standards Project: A Conversation with Lauren Resnick and Warren Simmons,” *Education Leadership* 50:5 (February, 1993): 17—21; and Rothman, *Measuring Up*, 123—128.

⁹¹ Linda Darling-Hammond, Jacqueline Aneess, and Beverly Falk, *Authentic Assessment in Action: Studies of Schools and Students at Work* (New York: Teachers College Press, 1995); *Authentic Reading Assessment: Practices and Possibilities* ed. Sheila W. Valencia (Newark: DE, International Reading Association, 1994).

⁹² Edmund W. Gordon, *Education and Justice: A View from the Back of the Bus* (New York: Teachers College Press, 1999), chapter 8.

⁹³ Edward Haertel, “Performance Assessment and Education reform,” *Phi Delta Kappan* 80:9 (May, 1999): 662—666. For a sketch of New Standards Project, see Rothman *Measuring Up*, pp. 123—127. A more thorough account of the history of the New Standards Project is forthcoming in a paper commissioned by the Gordon Commission on the Future of Assessment in Education. On the National Council on Education Standards and Testing, see Rothman, *Measuring Up*, 178—9.

⁹⁴ Diana Rhoten, Martin Carnoy, Melissa Chabren, and Richard Elmore, “The Conditions and Characteristics of Assessment and Accountability,” in *The New Accountability: High School and High-Stakes Testing*, ed Martin Carnoy, Richard Elmore and Leslie Santee Siskin (New York: Routledge-Falmer, 2003): 38—40.

⁹⁵ Editorial, *The Daily Gazette* (Schenectady, NY, November 22, 2001): B14.

⁹⁶ Shepard, “A Brief History of Accountability Testing,” 33.

⁹⁷ Cohen and Moffitt, *Ordeal of Equality*. On the Clinton administration’s education programs and the politics they generated, see Vinovskis, *From A Nation at Risk*, 67—109, 1224—131, 140—152; Jennings, *Why National Standards and Tests?*; Elizabeth H. DeBray, *Politics, Ideology, & Education: Federal Policy During the Linton and Bush Administrations* (New York: Teachers College Press, 2006) Chapters 3—5; and Cross, *Political Education*, 104—122.

⁹⁸ Claire Needell Hollander, “Teach the Books, Touch the Heart,” *New York Times* (April 20, 2012).

⁹⁹ Rothman, *Something in Common*, is well written and gives much information from the point of view of a participant. For more information and up-to-date figures on

participation, see the Common Core home site on line at www.corestandards.org accessed August 9, 2012, when 45 states had joined.

¹⁰⁰ See Catherine Gewertz, “Two Versions of ‘Common’ Test Eyed by State Consortium,” *Education Week* 32: 4 (September 19, 2012): 1, 19 and Catherine Gewert, “Testing Group Scales Back Performance Items,” *Education Week* (November 29, 2012) on line at <http://www.edweek.org/ew/articles/2012/11/30/13tests.h.32.html>, accessed on January 7, 2013.

¹⁰¹ I have discussed theories on this matter in “Clio at the Table: Historical Perspectives and Policymaking in the Field of Education,” in *Clio at the Table: Using History to Inform and Improve Education Policy* ed. Kenneth Wong and Robert Rothman (New York: Peter Lang, 2009), 283—294.

¹⁰² See Carl F. Kaestle, “Literate America: High-level Adult Literacy as a National Goal” In *Historical Perspectives on the Current Education Reforms*, eds., Diane Ravitch and Maris Vinovskis. (Baltimore: Johns Hopkins University Press, 1995): 329-354.

¹⁰³ Richard Herrnstein and Charles Murray, *The Bell Curve: Intelligence and Class Structure in America* (New York: The Free Press, 1994); for critiques, see James J. Heckman, “Lessons from the Bell Curve,” *Journal of Political Economy* 103:5 (1995): 1091—1120; and *The Bell Curve Wars: Race, Intelligence and the Future of America* ed. Steven Fraser (New York: Basic Books, 1995).

¹⁰⁴ *How People Learn: Brain, Mind, Experience, and School*, ed. John D. Bransford, Ann L. Brown, and Rodney R. Cocking (Washington, D.C.: National Research Council, Committee on Developments in the Science of Learning, National Academy Press, 1999); *Knowing What Students Know: The Science and Design of Educational Assessment* ed. James W. Pellegrino, Naomi Chudowsky, and Robert Glaser (Washington, D.C.: National Research Council, Committee on the Foundations of Assessment, National Academy Press, 2001); Michael E. Martinez, *Education as the Cultivation of Intelligence* (Mahrah, NJ: Lawrence Erlbaum Associates, 2000, e-book edition, Taylor & Francis, 2009).

¹⁰⁵ Lauren B. Resnick and Larry Berger, “An American examination System” (paper, National Conference on Next Generation K—12 Assessment Systems, March, 2010) on line at <http://tiny.cc/zte1aw>, accessed on August 1, 2012.