CHAPTER 4

Lessons From the Past: A History of Educational Testing in the United States

Contents

	Page
Highlights	103
Achievement Tests Come to American Schools: 1840 to 1875	104
Overview	104
Demography, Geography, and Bureaucracy	104
The Logic of Testing	108
Effects of Test Use	109
Science in the Service of Management: 1875 to 1918	110
Issues of Equity and Efficiency	111
An Intellectual Bridge	111
Mental Testing	112
Testing in Context	114
Managerial Efficiency	115
Achievement and Ability Vie for Acceptability	116
Experimentation and Practice	118
The Influence of Colleges	118
World War I	119
Testing Through World War II: 1918 to 1945	120
Overview	120
A Legacy of the Great War	121
The Iowa Program	122
Multiple Choice: Dawn of an Era	124
Critical Questions	124
College Admissions Standards: Pressure Mounts	125
Testing and Survey Research	126
Testing and World War II	127
Equality, Fairness, and Technological Competitiveness: 1945 to 1969	127
Overview	127
Access Expands	128
Developments in Technology	129
Race and Educational Opportunity	130
Recapitulation	131

Box

Box	
4-A. Mental Testers: Different Views	113

Figure

4-1. Annual Immigration to the United States: 1820-60 105

CHAPTER 4

Lessons From the Past: A History of Educational Testing in the United States

Highlights

- Since their earliest administration in the mid-19th century, standardized tests have been used to assess student learning, hold schools accountable for results, and allocate educational opportunities to students.
- Throughout the history of educational testing, advances in test design and innovations in scanning and scoring technologies helped make group-administered testing of masses of students more efficient and reliable.
- High-stakes testing is not a new phenomenon. From the outset, standardized tests were used as an instrument of school reform and as a prod for student learning.
- Formal written testing began to replace oral examinations at about the same time that American schools changed their mission from servicing the elites to educating the masses. Since then tests have remained a symbol of the American commitment to mass education, both for their perceived objectivity and for their undeniable efficiency.
- Although standardized tests were seen by some as instruments of fairness and scientific rigor applied to education, they were soon put to uses that exceeded the technical limits of their design. A review of the history of achievement testing reveals that the rationales for standardized tests and the controversies surrounding test use are as old as testing itself.

The burgeoning use of tests during the past two decades—to measure student progress, hold students and their schools accountable, and more generally solidify various efforts to improve schooling-has signified to some observers a ". . . profound change in the nature and use of testing. . . "[1] But the use of tests for the dual purposes of measuring and influencing student achievement is not a historical anomaly. The three principal rationales for student testing--classroom feedback; system monitoring; and selection, placement, and certification-have their roots in practices that began in the United States more than 150 years ago. And many of the points that frame the testing debate today, such as the potential for test misuse, echo arguments that have been sounded since the beginning of standardized student testing.

This chapter surveys the evolution of student testing in American schools, and develops four themes:

1. Tests in the United States have always been used to ascertain the effects of schooling on children, as well as to manage school systems and influence curriculum and pedagogy. Tests designed and administered from beyond classrooms have always been more useful to administrators, legislators, and other school authorities than to classroom teachers or students, and have often been most eagerly applied by those seeking school reform.
2. The historical use of standardized tests in the United States reflects two fundamentally American beliefs about the organization and allocation of educational opportunities: fairness and efficiency. The fairness principle involves, for example, assurances to parents that their children are offered opportunities similar to those given children in other schools or neighborhoods. Efficiency refers to the orderly provision of educational services to all children. These have been the foundation blocks for the

[1] George Madaus, quoted in Edward B. Fiske, "America's Test Mania," The New York Apr. 10, 1988, section 12, p. 18. See ch. 3 of this report for a detailed account of the rise of testing in the 1970s and 1980s.

297-933 0 - 92 - 8 QL 3

American system of mass public schooling; testing has been a key ingredient of the mortar.

3. Increased testing has engendered tension and controversy over its effects. These tensions reflect the centrality of schooling in American life, and competing visions of the purposes and methods of education within American pluralism. *Demand* for tests stems in large part from demand for fair treatment of all students; the use of tests, however, especially for sorting and credentialing of young persons, has always raised its own questions of fairness.

4, As long as schooling continues to play a central role in American life, and as long as tests are used to assess the quality of education, testing will occupy a prominent place on the public policy agenda. The search for better assessment technologies will continue to be fraught with controversies that have as much to do with testing per se as with conflicting visions of American ideals and values.

This chapter focuses on testing through four chronological periods. The first section begins with the initial educational uses of standardized written examinations in the mid-19th century and continues through the development of mental (intelligence) measurement near the end of that century. The next section covers the onset of intelligence and achievement testing in the schools, a movement spurred largely by managerial and administrative concerns and supplied, in large part, with the newly developing tools of ''scientific' testing. The third section focuses on trends in educational testing from the end of World War I through the end of World War II, a period marked by important technological advances as well as refinements in the art and science of testing. The last section of this chapter is a discussion of the pivotal role of testing in the struggle for racial equality, increased educational access, and international technological competitiveness in the years after World War II.

Achievement Tests Come to American Schools: 1840 to 1875

Overview

The period from 1840 to 1875 established several main currents in the history of American educational testing. First, formal written testing began to replace oral examinations administered by teachers and schools at roughly the same time as schools changed their mission from servicing the elite to educating the masses. Second, although the early standardized examinations were not designed to make valid comparisons among children and their schools, they were quickly used for that purpose. Motivated in part by a deep commitment to fairness in educational opportunities, the use of tests soon became controversial precisely over challenges to their fairness as a basis for certain types of comparisons-challenges leveled by some teachers and school leaders, although not by the most active crusaders on behalf of free and universal education. Third, the early written examinations focused on the basics-the major school subject--even though the objectives of schooling were understood to be considerably broader than these topics. Finally, from their inception standardized tests were perceived as instruments of reform:[2] it was taken as an article of faith that test-based information could inject the needed adrenalin into a rapidly bureaucratizing school system.

Demography, Geography, and Bureaucracy

Tests of achievement have always been part of the experience of American school children. In the colonial period, school supervisors administered oral examinations to verify that children were learning the prescribed material. Later, as school systems grew in size and complexity, the design, purposes, and administration of achievement testing evolved in an effort to meet new demands. Well before the Civil War, schools used externally mandated written examinations to assess student progress in specific curricular areas and to aid in a

[2] ''Reform'' means different things to different people, especially with respect to education. In this report, the word is intended neutrally, i.e., as "change," although it clearly connotes the intention to improve, upgrade, or widen children's educational experiences. The possibility that good intentions can lead to unintended consequences is the central theme in such works as Michael B. Katz, *Irony of Early Reform* (Cambridge, MA: Harvard University Press, 1968). See also Lawrence Cremin, *Transformation of Progressivism* York, NY· Vintage Books, 1964) for an even broader exploration of change, i.e., as "transformation" of the school.

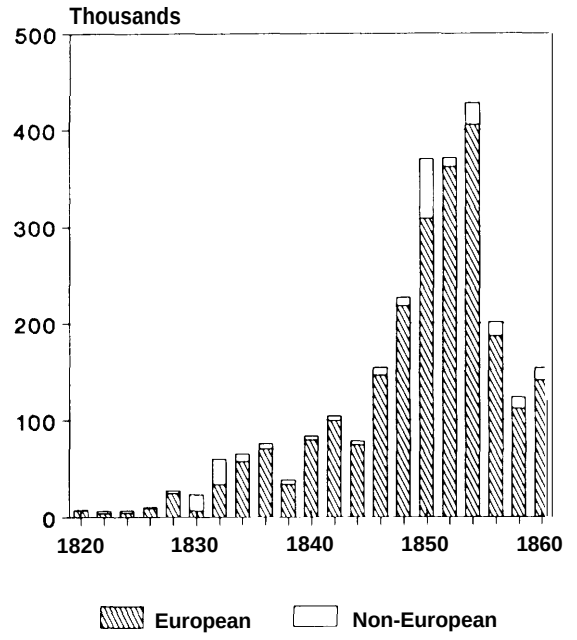
variety of administrative and policy decisions.[3] As early as 1838 American educators began articulating ideas that would soon be translated into the formal assessment of student achievement.

What were the main factors that led to this interest in testing? What were the main purposes for testing? Some of the answers lie in the demography and political philosophy that shaped the 19th century American experience.

Between 1820 and 1860 American cities grew at a faster rate than in any other period in U.S. history, as the number of cities with a population of over 5,000 increased from 23 to 145.[4] That same period saw an average annual immigration of roughly 125,000 newcomers, mostly Europeans (see figure 4-1).[5] Coincident with this immigration and urbanization, the idea of universal schooling took hold. By 1860 ". . . a majority of the States had established public [primary] school systems, and a good half of the nation's children were already getting some formal education."[6] Some States, like Massachusetts, New York, and Pennsylvania, were moving toward free secondary school as well.

Although it is difficult to establish a causal link between these demographic and educational changes, surely one thing that attracted European immigrants was the ideal of opportunity embodied in the American approach to universal schooling. Following his visit to the United States in 1831 to 1832, the Frenchman Alexis de Tocqueville shared with his countrymen his conviction that there was no other country in the world where ''. . . in proportion to the population there are so few ignorant and at the same time so few learned individuals. Primary instruction is within the reach of everybody; superior instruction is scarcely obtained by any."[7]

Figure 4-1—Annual Immigration to the United States: 1820-60



SOURCE: Office of Technology Assessment, based on data from U.S. Department of Commerce, Bureau of the Census, *Historical of the United* (Washington, DC: 1975), pp. 105-111.

At the same time, it could be argued that population growth and increased heterogeneity necessitated the crafting of institutions-such as universal schooling-to "Americanize" the masses. The 20th century social philosopher Hanah Arendt wrote, for example, that education has played a "... different, and politically incomparably more important, role [in America] than in other countries,' in large part because of the need to Americanize the immigrants.[8]

The concept of Americanization extended well beyond the influx of immigrants who arrived in the latter half of the 19th century, however. The

[3]Many historians of American educational testing focus on the influence of the intelligence testing movement, which began at the end Of the 19th century. See, e.g., Daniel Resnick, "The History of Educational Testing," *Ability* part 2, Alexandra Wigdor and W. Garner (eds.) (Washington DC: National Academy Press, 1982), pp. 173-194; or Walter Haney, "Testing Reasoning and Reasoning About Testing, " vol. 54, No. 4, winter 1984, pp. 597-654.

[4]David Tyack, *The One Best System: A History* (Cambridge, MA: Harvard University Press, 1974), p. 30.

[5]U.S. Department of Commerce, Bureau of the Census, *Historical Statistics of the United States: Colonial Times 1970*, part 1 (Washington, DC: U.S. Government Printing Office, 1975), p. 106.

[6]Cremin, op. cit., footnote 2, p. 13. This chapter relies heavily on Cremin's work, but also on important educational historiography of David Tyack, Michael Katz, Ira Katznelson, Margaret Weir, and Carl Kaestle.

[7]See Alexis de Tocqueville, *Democracy in America*, vol. 1 (New York, NY: Vintage Books, July 1990), p. 52.

[8]Hannah Arendt, "The Crisis in Education," *Partisan Review*, vol. 25, No. 4, fall 1958, pp. 494-495. See also Diane Ravitch, *Great School Wars:* 1805-1973 (New York, NY: Basic Books, 1974), p. 171, for her treatment of some of the early American educators (like William Henry Maxwell in New York) who saw schooling as the ". . . antidote to problems that were social, economic, and political in nature. "

foundation for a political role for education had already been laid in the colonial and post-Revolutionary periods, as religious, educational, and civic leaders began considering the possible relationships between lack of schooling, ignorance, and moral delinquency. These leaders, especially in the burgeoning cities, advocated public schooling for poor children who lacked access to church-run charity schools or to common pay schools (schools available to all children in an area but for which parents paid part of the instructional costs).

Up until the mid-19th century, the pattern of education consisted of private schools run by paid tutors, State-chartered academies and colleges with more formal programs of instruction, benevolent societies, and church-run charity schools—in sum, a "hedge-podge' reflecting the many:

> . . . motives that impelled Americans to found schools: the desire to spread the faith, to retain the faithful, to maintain ethnic boundaries, to protect a privileged class position, to succor the helpless, to boost the community or sell town lots, to train workers or craftsmen, to enhance the virtue or marriageability of daughters, to make money, even to share the joys of learning.[9]

Population growth and density created new strains on schools' capacity to provide mass education.[10] According to census statistics, public school enrollments grew from 6.8 million in 1870 to 15.5 million by 1900. By the turn of the century, almost 80 percent of children aged 5 to 17 were enrolled in some kind of school.[11] Mass public education could no longer be viable without fundamental institutional adaptations. Expanding enrollments also placed new strains on the public till as public school began overshadowing private and charity schools. In direct expenditures, the percentage of total education spending attributable to the public schools grew from less than one-half in 1850 to more than 80 percent in 1900.[12] In terms of foregone income as well, the costs were impressive: the income that students aged 10 to 15 would have earned were they not in school increased from an estimated nearly $25 million in 1860 to almost $215 million in 1900.[13] Not surprisingly, this spending inevitably led to calls for evidence that the money was being used wisely.

The size and concentration of the growing student population increased the taxpayers' burden and created new institutional demands for efficiency similar to those that governed the evolving nature of many American institutions. One way schools could demonstrate sound fiscal practice was by organizing themselves according to principles of bureaucratic management. "Crucial to educational bureaucracy was the *objective and efficient classification,* or grading, of pupils."[14] According to Henry Barnard, a prominent figure in the common school movement, it was not only inefficient, but also inhumane, to fill a classroom with children of widely varying ages and *attainment.*[15] *On this* assumption, the mid-19th century reformers sought additional information that would make the classification more rational and efficient than the prevailing system of classification, based primarily on age. They turned their attention toward achievement tests.

The result was one of many ironies in the history of educational testing: the classification and grouping of students, essentially a Prussian idea, became a pillar in the public school movement that was an American creation. No less an American educational statesman than Horace Mann, who saw universal

---

[9]David Tyack and Elisabeth Hansot, (New York, NY: Basic Books, 1982), p. 30. See also Katz, op. cit., footnote 2, p. 131. Katz writes that: ". . . the duty of the school was to supply that inner set of restraints upon passion, that bloodless adherence to a personal sense of rights, which would counteract and so reform the dominant tone of society. "

[10]For a more detailed *analysis* of the shifts from rural to urban education, see, e.g., Tyack, Op. Cit., footnote 4. Also, see Michael B. Katz, *Class,* York, NY: Praeger, 1972).

[11]Bureau of the Census, op. cit., footnote 5, p. 369. See also Tyack, op. cit., footnote 4, p. 66, who cites a report by W.T. Harris with similar data.

[12]Tyack and Hansot, Op. Cit., footnote 9, p. 30.

[13]Tyack, op. cit., footnote 4, pp. 66-67.

[14]Ibid., p. 44, emphasis added. It is worth recalling that the early exponents of bureaucracy spoke Of its formalism—manifest in classification systems of the type discussed here-in positive terms, i.e., as an improvement over earlier forms of organization that were at once less fair and less efficient. See, e.g., Max Weber, *of Social Organization,* edited and translated by A.M. Henderson and T. Parsons (New York, NY: Macmillan Publishing Co., 1947). The appeal of tests as both fair and efficient tools of management is a main theme in this chapter.

[15]Tyack, *Op.* cit., footnote 4, p. 44, emphasis added. B arnard's lifelong commitment to school improvement for the masses, coupled with his belief in the importance of conserving the social and economic status of the privileged classes, personifies an important aspect of the American experiment with democratic education. See also Merle Curd, *Social Ideas* (Paterson, NJ: Pageant Books, Inc., 1959), pp. 139-168.

Photo credits: Frances B. Johnston

Teachers have always assessed student performance directly, These photos were taken circa 1899 for a survey of Washington, DC schools.

education as the ''great equalizer' and who had a ... total faith in the power of education to shape the destiny of the young republic,"[16] supported the highly structured model of schools in which students would be sorted according to their tested proficiency.[17] Thus, as early as the mid-19th century, there existed a belief in the role of testing as a vehicle to classify students *ex ante,* commonly viewed as a necessary step in providing education. Also emerging during this period was an interest in uses of tests *ex post:* to monitor the effectiveness of schools in accomplishing their purposes. Visionaries like Mann saw testing as a means to educate effectively; administrators, legislators, and the general public turned to tests to see what children were actually learning.

In fact, it was during Horace Mann's tenure as Secretary of the (State) Board of Education that Massachusetts became the site of ". . . the first reported use of a written examination . . . ***after some harassment by the State Superintendent of Instruction about the shortcomings of the schools. . .*** "[18] From its inception, this formal written testing had two purposes: to classify children (in pursuit of more efficient learning)[19] and to monitor school systems by external authorities. Under Mann's guidance, the State of Massachusetts moved from subjective oral examinations to more standardized and objective written ones, largely for reasons of efficiency. Written tests were easier to administer and offered a streamlined means of classifying growing numbers of students.

[16]Cremin, op. cit., footnote 2, pp. 8-9.

[17]Katz, op. cit., footnote 2, pp. 139-140.

[18]Resnick, op. cit., footnote 3, p. 179, emphasis added.

[19]Tyack, op. cit., footnote 4, p. 45. Tyack notes that classification preceded standard examinations: ". . . the proper classification was only the beginning. In order to make the one best system work, the schoolmen also had to design a uniform course of study and standard examinations." But he does not describe the criteria for classification used prior to the standard examinations, which would be important to analyze the comparative fairness of formal and informal classification systems. It appears, though, always to have involved some type of proficiency testing, the difference being between the looser and more subjective classroom-based tests and the more format externally administered tests.

It is important to point out what "standardization" meant in those days. It did not mean "norm-referenced" but rather that ". . . the tests were published, that directions were given for administration, that the exam could be answered in consistent and easily graded ways, and that there would be instructions on the interpretation of results. ' '[20] The model was quite consistent with the assumed virtues of bureaucratic management. The efficient flow of information was not unique to education or educational testing; it was becoming a ubiquitous feature of American society .21

Perhaps more important, though, was the evolving role of testing as a vehicle to ensure fairness and evenhandedness in the distribution of educational resources: *one way to ascertain whether children in the one-room rural schoolhouse were receiving the same quality of education as their counterparts in the big cities was to evaluate their learning through the same examinations.* Thus, standardized testing came to serve an important symbolic function in American schools, a sort of technological embodiment of principles of fairness and universal access that have always distinguished American schools from their European and Asian counterparts. As the methods of testing later became increasingly quantitative and "scientific" in appearance, the tests gained from the growing public faith in the ability of science and rational decisionmaking to better mankind.

But Mann had other reasons for introducing standardized testing. He had been engaged in an ideological battle with the Boston headmasters, who perceived him as a "radical." This disagreement reflected a wider schism in the Nation between reformers like Mann who believed in stimulating student interest in learning through greater emphasis on the "real world,' and hard-liners who believed in discipline, rote recitation, and adherence to texts.[22] Although Mann and his compatriots eventually won, setting American public education on a unique historical course, one of their more potent weapons in the battle was one that might today be associated with a hard-line, top-down approach to school reform: when two of Mann's allies were appointed to examine the status of the grammar schools, ". . . they gave written examinations with questions previously unknown to the teachers [and] . . . published a scathing indictment of the Boston grammar schools in their annual report. . . ."[23]

The Logic of Testing

The fact that the first formal written examinations in the United States were intended as devices for sorting and classifying but were used also to monitor school effectiveness suggests how far back in American history one can go for evidence of test misuse. The ways in which these tests were used for monitoring was logical: to find out how students and their schools are performing, it made sense to conduct some sort of external measurement process. But the motivation for the standardized examinations in Massachusetts was, in fact, more complicated and reveals a pattern that would become increasingly familiar. *The idea underlying the implementation of written examinations, that they could provide information about student learning, was born in the minds of individuals already convinced that education was substandard in quality. This* sequence-perception of failure followed by the collection of data designed to document failure (or success)--offers early evidence of what has become a tradition of school reform and a truism of student testing: tests are often administered not just to discover how well schools or kids are doing, but rather to obtain external *confirmation-validation*— of the hypothesis that they are not doing well at all.[24]

---

[20] Resnick, op. cit., footnote 3, p. 179.

21 George Madaus, *for* example, writes that the movement toward standardization and conformity began in 1815 with efforts in the Army Ordnance Department to develop ". . . administrative, communication inspection accounting, bureaucratic and mechanical techniques that fostered conformity and resulted in the technology of interchangeable parts . . . [and that] these techniques . . . were well known throughout the textile mills and machines shops of New England when Horace Mann introduced the standardized written test. . . .' George Madaus, '"lksting as a Social Technology,' unpublished monograph Inaugural Annual Boisi Lecture on Education and Public Policy, Boston College, Dec. 6, 1990, pp. 26-27. See also Katz, op. cit., footnote 2, pp. 5-11, for an account of the dramatic changes in the structure and management of American business during Mann's lifetime.

[22] See Katz, op. cit., footnote 10, pp. 115-153, for a fuller discussion of the origins and ramifications of this ideological struggle.

[23] Ibid., p. 152. See also Madaus, Op. Cit., footnote 21.

[24] Although testing was not yet considered a scientific enterprise (that would come later in the century, with the emergence of psychology and the concepts of mental measurement-see below), the logic of its application had traces of the inductive model: from empirical observations of the schools, to hypotheses explaining those observations, to the more systematic and less anecdotal collection of data in order to test the hypotheses. For a physicist's views on the basic fallacies in mental "measurement,' however, see David Layzer, "Science or Superstition? A Physical Scientist Looks at the IQ Controversy," *IQ Critical Readings*, N.J. Block and Gerald Dworkin (eds.) (New York, NY: Pantheon Books, 1976), pp. 194-241.

The use of formal, written achievement tests in Massachusetts (and soon afterwards in many other places), as already emphasized, was motivated largely by administrative concerns.[25] The tests themselves often focused on a rather narrow set of outcomes, selected principally to put the headmasters in the worst possible light. There was a profound mismatch between the content covered in those early achievement tests and the objectives of common schooling those tests were intended to gauge. Given the schools' broad democratic agenda, and given the environment of demographic and geographic shift in which the agenda was to be carried out, the estimation of educational quality by a ". . . test of thirty questions on the subjects scheduled for study during the year. . . given to about half the eighth grade, one thousand students, "[26] is a telling early example of the limitations of tests in measuring the range of knowledge students acquire during a school year.

From their inception, written achievement tests were among the more potent weapons of reform of teaching and school administration. For example, Samuel Gridley Howe, an ally of Mann, looked to tests to provide ''. . . a single standard by which to *judge and compare the output of each school,* 'positive information in black and white,' [in place of] the intuitive and often superficial written evaluation of oral examinations."[27]

The tests Mann and Howe encouraged covered a narrow range of school material; there was no attempt to link students' test performance with specific features of school organization or pedagogy; and the schoolmasters usually selected which students took the tests.[28] But these technical issues did not interfere with the use of test results as a basis for reform. Mann, for one, successfully convinced his fellow Bostonians that the tests were able to "... determine, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught. '[29] Teachers, for their part, went along with the testing as long as they saw it as a way to wield power over their students.[30]

Effects of Test Use

Not surprisingly, soon after the frost application of tests came criticisms that have also become a steady presence in school life. First, there was public amazement at the poor showing of the test-takers: ''Out of 57,873 possible answers, students answered only 17,216 correctly and accumulated 35,947 errors in punctuation in the process. Bloopers abounded: one child said that rivers in North Carolina and Tennessee run in opposite directions because of 'the will of God.'" Second, it was feared that the tests were driving students to learn by rote: . . . [according to Howe] they could give the date of the embargo but not explain what it did."[31]

Nevertheless, test use continued, and from the earliest applications, test use raised key questions. Consider, for example, that the main beneficiaries of test information were not the teachers and principals, who might have used it to change aspects of their specific institutions, but rather State-level policy - makers and administrators. Thus, while there might have been a casual acceptance of the principle that tests could provide information necessary to effect change, there was apparently much less agreement—or perhaps just simple naivete-as to how and where the changes would be initiated. 'The most important reported result, *an unintended one from the standpoint of the [Boston] school committee,* was to make city teachers and principals accountable to supervisory authority at the State level. ' '[32] Tests became

[25] Schools were not alone in their growing admiration for quantification. Prison reformers, abolitionists, and others were *also* fond of statistics. For a lucid discussion of the reverence for science and quantitative methods, which would peak at the turn of the century, see Paula S. Fass, ''The IQ: A Cultural and Historical Framework,'' *Education,* vol. August 1980, pp. 431-458.

[26] Resnick, op. cit., footnote 3, p. 179.

[27] Tyack, op. cit., footnote 4, p. 35, emphasis added.

[28] ''Even within the grade, [the Boston test] was not a fair sample of students, since the schoolmasters were free to choose who would take the test. " Resnick, op. cit., footnote 3, p. 179.

[29] Quoted in Paul Chapman, *Schools as Sorters: Lewis Terman, 1890-1939* (New York University Press, 1988), p. 33.

[30] Robert Hample, University of Delaware, personal communication May 1991.

[31] Tyack, op. cit., footnote 4, p. 35. According to Tyack, Howe knew how ''abstruse and tricky' the test items were, but thought it was a fair basis for comparison of students nonetheless. Given the reference to punctuation errors, it seems that the tests included at least some written work; in any event, we know that multiple choice was not invented until several decades later, which suggests that test format is not the sole' determinant of content validity, fairness, or the tendency to learn.

[32] Resnick, *Op.* cit., footnote p. 180, emphasis added.

important tools for education policymakers, despite their apparently limited value to teachers, students, and principals.

A related development offers yet another illustration that current problems in educational testing are not all new. Although the written examinations were intended to provide information about schools and students, that information was not necessarily meant to become a basis for *comparisons*. Yet that is quickly what happened, as illustrated in the case of examinations used for high school admission: "Although only a minority of students took the [standard short-answer] exam, performance [on the exam] . . . could function, within the larger communities, to compare the performance of classes from different feeder schools."[33]

The case cited in this example points to a pervasive dilemma in the intended and actual uses of tests. On the one hand, information about student performance was understood to be essential as a basis for organizing classroom learning and judging its output; on the other hand, once the information was created, it was quickly appropriated to uses for which it had not been designed-specifically, to comparisons among schools and districts. The fact that the jurisdictions were different in so many fundamental ways as to render the comparisons virtually meaningless did not seem to matter. Nevertheless, by the 1870s many school leaders were beginning to question the comparisons: ''. . . a careful observation of this practice for years has convinced me that such comparisons are usually unjust and mischievous. ' '[34] At the same time, there was widespread agreement that ''. . . the classroom was part of the production line of the school factory [and that] examinations were the means of judging the value added to the *raw material* . . . during the course of the year."[35]

In the latter part of the 19th and early 20th centuries, changing demography would continue to influence school and test policy. Other factors would also begin to play a role: the development of psychology and ''mental measurement' as a science, and the increasing influence of university and business interests on performance standards for the secondary schools. These are the main topics in the next section of the chapter.

Science in the Service of Management: 1875 to 1918

During the period from 1875 to the end of World War I, the development and administration of a range of new testing instruments-from those that sought to measure mental ability to those that attempted to assess how well students were prepared for college-brought to the forefront several critical issues related not only to testing but to the broader goals of American education. First, as instruments that were designed to discern differences in individual intelligence became available, the concept of classifying and placing students by ability gained greater acceptance, even among those who espoused the democratic ideals of fairness and individuality.

Second, as research on mental measurement continued, it gave rise to new debates about the role of heredity in determining intellectual ability and the effects of education. Some theorists used the results of intelligence and aptitude tests to support claims of natural hierarchy and of racial and ethnic superiority.

Third, mirroring the structural changes occurring in businesses and other American institutions, school systems reorganized around the prevailing principles of efficient management: consolidation of small schools and districts, classification of students, bureaucratization of administrative responsibilities. Within these new arrangements, tests were viewed as an important efficiency tool.

Fourth, by the end of World War I, standardized achievement tests were available in a variety of basic subjects, and the possibilities for large-scale group testing had been demonstrated. The results of these tests gave reformers (including college presidents) ammunition in their push for improvements in educational quality.

Fifth, the implementation of mass testing in World War I ushered in a new era of educational testing as well.

[33] Ibid. For an indepth study of the role of tests and other criteria in admissions decisions at Philadelphia's Central High School, see David F. Labaree, Haven, CT: Yale University Press, 1988), especially chs. 3 and 4.

[34] Emerson White (an early leader in the National Education Association), quoted in Tyack, Op. cit., footnote 4, P. 49.

[35] John Philbrick, quoted in ibid., p. 49, emphasis added.

Issues of Equity and Efficiency

The analysis in the preceding section of this chapter raises a perplexing question about the role of testing in American education: how could the emerging American and democratic theory of education be reconciled with standardized tests that covered, at best, a small portion of what schooling was supposed to accomplish, and, at worst, were used in ways that violated basic democratic principles of fairness? Part of the answer in the early years of testing lay in the role of curriculum in the public school philosophy. Horace Mann, for example, was

> ... inclined to accept the usual list of reading, writing, spelling, arithmetic, English grammar, and geography, with the addition of health education, vocal music (singing would strengthen the lungs and thereby prevent consumption), and some Bible reading. "[36]

Thus, it might be argued that one reason Mann favored the formal examinations was that they signaled the importance of learning the major subjects, which, in his view, was the first step toward achieving the broader goals of morality, citizenship, and leadership. Learning the major subjects was a necessary-if insufficient-condition for education writ large.[37]

Another factor was that because standardized tests were new, there was no established methodology for designing them or judging whether test scores accurately reflected learning. Furthermore, school reformers seemed relatively unconcerned that emphasizing the basics might compromise the broader objectives of schooling. Generally they viewed the basics as just that: the necessary building blocks on which the broader objectives of education could be erected.

If that explanation helps resolve the curious acceptability of short tests as proxies for complex educational goals, it does not offer any obvious clues to the paradox that the use of tests to track students had its roots in the movement to universalize and democratize education. Again, Mann's thinking on the subject can shed some light. Although "Mann was one of the first after Rousseau to argue that education in groups is not merely a practical necessity but a social desideratum, ' '[38] he had an equally powerful belief in individuality. Mann's answer was to tailor lessons in the classroom to meet the needs of individual children: ". . . children differ in temperament, *ability,* and interest . . " and need to be treated accordingly .[39] From here, then, it was not a far leap to embracing methods that, because they were purported to measure those differences, could be used to classify children and get on with the educational mission.

Mann was not alone. The American pursuit of efficiency would become the hallmark of a generation of educationists, and would create the world's most fertile ground for the cultivation of educational tests.

An Intellectual Bridge

Some social scientists have characterized mental measurement-a branch of psychology that blossomed during the late 19th and early 20th centuries and prefigured modern psychological testing-as

> ... the most important single contribution of psychology to the practical guidance of human affairs."[40]

Psychological testing was able to flourish because of its appeal to individuals of nearly every ideological stripe. It was not just the hereditarians and eugenicists who were attracted to such concepts as ''intelligence' and the ''measurement' of mental ability; many of the early ***believers in the*** measurement of mental and psychophysical processes were progressives, egalitarians, and communitarians committed to the betterment of all mankind.

Mann, for one, embraced phrenology-an approach to the assessment of various cognitive capacities based on physical measurement of the size of areas of the brain-without reservation, joining the ranks of such advocates as Ralph Waldo Emerson, Walt Whitman, William Ellery Charming, Charles Sumner, and Henry Ward Beecher, as well

---

[36]Cremin, op. cit., footnote 2, p. 10.

[37]The belief that learned persons were better, in the moral sense, has been pervasive throughout the history of American education, See, e.g., Curti, op. cit., footnote 15. A major figure in the measurement of ability and achievement, Edward Thorndike, produced empirical results showing the high correlation between intellectual attainment and morality. See, e.g., Tyack and Hansot, op. cit., footnote 9, p. 156.

[38]Cremin, op. cit., footnote 2, p. 11.

[39]Ibid.

[40]Lee Cronbach, "Five Decades of Public Controversy Over Mental Testing,'' January 1975.

as a host of respected physicians.[41] Phrenology attributed good or base character traits to differences in physical endowments; Mann and others saw in this doctrine a persuasive rationale for education as a means of cultivating every individual's admirable *propensities* and checking his coarser ones. One might say, then, that phrenology symbolized to Mann a unique chance to mobilize support for social intervention.[42]

Phrenology was a methodological bridge from crude comparisons based on written achievement examinations, to measures that were at once more scientifically rigorous and more sensitive to innate differences in ability.[43] The principal intelligence researchers whose work would ultimately be translated into the American science of mental testing—Galton, Wundt, and Binet--had each dabbled in phrenology before devising their methods for assessing human intelligence.

Mental Testing

In the late 19th century, European and American psychologists began independently seeking ways to corroborate and measure individual differences in mental ability. Sir Francis Galton in England and J. McKeen Cattell in the United States conducted a series of studies-mostly dealing with sense perception but some focusing on intellectual aptitude--that may be said to mark the beginning of modern intelligence testing.[44] It was Cattell, in fact, who coined the term "mental test" in a paper published in 1890.

In an effort to trace the hereditary origins of mental differences, Galton conducted the first empirical studies of the heritability of mental aptitude and developed the first mental test, although he did not call it that.[45] Although the more extreme views of some of these early researchers have long since been repudiated, and although some veered off into distasteful and unsupportable conclusions about hereditary differences (see box 4-A), their work nevertheless stimulated interest in intelligence testing that persists today.

The French psychologist and neurologist Alfred Binet also had a very strong influence on the development of intelligence tests in America and on their uses in schools, although not necessarily in the ways Binet himself would have liked. Empirically based definitions of intelligence and accounting explicitly for age were two of Binet's most important contributions to the science of mental testing. For Binet "intelligence" was not a measurable trait in and of itself, like height or weight; rather, it was only meaningful when tied to specific observable behaviors. But what behaviors to observe? Answering this question led Binet to his second major insight: ability to perform various mental behaviors varied with the age of the individual being observed. His research, therefore, consisted of giving children of different ages sets of tasks to perform; from their performances he computed average abilities-for those tasks-and how individual children compared on those tasks.[46] Neither the concept that intelligence existed as a unitary trait, nor the concept that individuals have it in freed amounts from birth, are attributable to Binet. Moreover, to Binet and coworker Theodore Simon, intelligence meant ''. . . judgment, otherwise called good sense, practical

41 About Mann's attraction to phrenology, historian Lawrence Cremin wrote: "It reached for naturalistic explanation of human behavior; it stimulated much needed interest in the problem of child health; and it promised that education could build the good society by improving the character of individual children. What a wonderful psychology for an educational reformer!'' Op. cit., footnote 2, p. 12.

42 Curti, op. cit., footnote 15, pp. 110-111. Michael Katz points out that "· -. to Mann and others of his time [intelligence] meant . . . a capacity that could be developed, not an innate limit on potential ., . an important point bemuse it shows that 'intelligence' is partly a social/cultural construction that we shouldn't reify. . . ." Personal communication, Aug. 18, 1991.

43 The history of phrenology contains some amusing ironies. Franz Gall, for example, one of the founders of the discipline, had to suffer the embarrassment of having his own brain weigh in ''at a meager 1,198 grams, " considerably lighter than the brains of real geniuses like Turgenev. For discussion see Stephen Jay Gould, *The Mismeasure of Man* (New York, NY: Norton, 1981), p. 92. And Francis Galton, whose own phrenologist surmised that his ". . . intellectual capacities are not distinguished by much spontaneous activity in relation to scholastic affairs. . . ." (Raymond E. Fancher, *Intelligence Makers IQ Controversy* York, NY: W.W. Norton & Co., 1985), p. 24), was later credited with launching the science of individual differences and of mental testing.

44 Walter S. Monroe, Ten *Years of Educational* *1918-1927* (Ural IL: University of Illinois, 1928), p. 89.

45 Borrowing methods of data collection and analysis from mathematics and astronomy, he also invented a statistical procedure that his student Karl Pearson would later turn into what is still the most powerful tool in the statistician's arsenal, the correlation coefficient.

46 Had the United States' move to universal public schooling begun in the late 19th century, and not in the middle, it is likely that the first achievement tests (described in the first section of this chapter) would have been more focused on innate ability and aptitude rather than on mastery of subjects taught in school. As will be shown below, however, the strands of ability and achievement ultimately did converge, largely due to the work of Terman and Thorndike.

Box 4-A—Mental Testers: Different Views

Although Charles Darwin himself never extrapolated from his biological and physical theory of evolution to evolution of cognitive abilities, Sir Francis Galton, his second cousin, made the leap. Galton's basic theory was that mental abilities were distributed unevenly in the population, and that while a certain amount of nurturing could have an effect, there was, as with physical ability, an upper bound predetermined by one's natural (genetic) endowments.

At the same time, researchers in the German laboratory of Wilhelm Wundt had also been involved in early studies of mental differences, with a focus on physical differences in sensation, perception, and reaction time. Apparently Wundt himself was not terribly interested in the development of tests for mental processes independent of the physical senses, but some of his students in the United States-such as Cattell--became prominent figures in the debate over hereditary origins of intelligence.

Although the name of Alfred Binet is commonly associated with the notion of IQ, Binet himself had strong reservations about using intelligence test data to classify and categorize children, and was opposed to the reduction of mental capacities to a single number. One reason had to do with his awareness of the difficulty of keeping the data purely objective. Another reason was his fear that "... individual children [would be] placed in different categories by different diagnosticians, using highly impressionistic diagnostic criteria . . . [and] . . . that the diagnosis was of particular moment in borderline cases."[1]

With his colleague Theodore Simon, Binet undertook an inductive study of children's intelligence: ". . . they identified groups of children who had been unequivocally diagnosed by teachers or doctors as mentally deficient or as normal, and then gave both groups a wide variety of different tests in their hopes of finding some that would differentiate between them."[2] Eventually they developed the key insight that the age of the child had to be considered in examining differences in test performance. The 1905 Binet/Simon test proved a workable model to make discriminations among the normal and subnormal populations of children. Binet, it should be noted, differed with many of his contemporaries on the role of heredity in intelligence. Binet believed that intelligence was fluid, ". . . shaped to a large extent by each person's environmental and cultural circumstances, and quantifiable only to a limited and tentative degree."[3]

Binet's followers took a different road than Binet himself would likely have chosen. Unlike those who worked in the tradition of Galton and who focused on measurement of young adults at the upper end of the ability distribution, Binet had devoted much of this part of his career to diagnosing retardation among children at the lower end of the distribution. And in fact, Binet's view of intelligence as a blend of multiple psychological capacities-attention, imagination, and memory among them-is enough to distinguish him from a generation of intelligence testers who followed, especially in the United States.

[1]Raymond E. Fancher, *The Intelligence Men: Makers Of the IQ* Controversy (New York, NY: W.W. Norton & Co., 1985), P- 70.

[2]Ibid., p. 70.

[3]Ibid., p. 82.

sense, initiative, the faculty of adapting one's self to circumstances. . . . A person may be a moron or an imbecile if he is lacking in judgment; but with good judgment he can never be either."[47] These characteristics of the Binet-Simon tradition were altered when the concepts of mental testing were imported to the United States.

Several Americans revised the Binet-Simon scale and adapted it for use in the United States. Stanford Professor Lewis Terman was perhaps the most influential and successful of the American mental testers. His 1912 revisions, called the Stanford Revision, caught on quickly and marked the beginning of large-scale individual intelligence testing in the United States.[48] As discussed in box 4-A, the *technology* of intelligence testing in the United States—in particular the connection between test performance and age in the formation of intelligence scales-was directly influenced by Binet; but the *philosophy underlying the use and interpretation of*

[47]A. Binet and T. Simon, *The Development of Intelligence in Children*, translated by E.S. Kite (Baltimore, MD: Williams and Wilkins, 1916), pp. 42-43. For discussion of the Binet-Simon tradition in intelligence testing, see, e.g., Robert Steinberg, *Metaphors Mind* (Cambridge, England: Cambridge University Press, 1990).

[48]Monroe, op. cit., footnote 44, p. 90.

the tests was inherited from Galton and his followers. Several historians have noted the mixed lineage of American testing; one has summarized it eloquently, noting that:

> . . . it was only as the French concern with personality and abnormality and the English preoccupation with individual and group differences, as measured in aggregates and norms, were superimposed on the older German emphasis on laboratory testing of specific functions that mental testing as an American science was born.[49]

Testing in Context

There is a tendency in the psychological literature to overstate the influence of Galton, Binet, and the other pioneers of mental testing on the demand for educational tests among American school authorities. That demand grew from a range of social and economic forces that produced similar calls for efficiency and compartmentalization in the workplace. Interest in the application of tests undoubtedly would have arisen even without the hereditarian influences of Galton and others who thought humankind could be bettered through gradual elimination of the subnormally intelligent.[50]

What was happening in the schools in the midst of these intellectual storms? For one thing, immigration was becoming an even more dominant influence on American political and social thinking. By 1890, some **15** percent of the American population was foreign born, and the quest for Americanization was continuing full steam. These "new' immigrants came from Southern and Eastern Europe (Austria, Hungary, Bulgaria, Italy, Poland, and Russia among others), and their numbers were beginning to overtake the traditional immigrants arriving from Northern Europe (Anglo-Saxons, French, Swiss, and Scandinavians). The effects on schools were staggering.

These abrupt demographic shifts affected many aspects of American life, but schools had a unique charge to maintain order in a society undergoing massive change and fragmentation and to inculcate American democratic values into massive numbers of immigrants. "Just as mass immigration was a symbol for-even the embodiment of-cultural disruption, education became its dialectical opposite, an instrument of order, or direction, of social consolidation. "[51] Because American schools were committed to principles of democratic education and universal access, instruments designed to bring order to schools without violating principles of fairness and equal access were extremely attractive.

Photo credit: Tamara Cymanski, OTA staff

Schools in America have played a central role in preparing immigrants for life in their new home. Challenged by the goals of educating massive numbers of newcomers fairly and efficiently, schools relied heavily on standardized testing.

Indeed, standardized tests offered even more than that. For one thing, they held promise as a tool for assessing the current condition of education, a means to gather the data from which reforms for integrating the masses could be designed. In what was perhaps the first effort to blend objective evaluation with journalistic-style muckraking, Joseph Mayer Rice conceived the idea of giving a uniform spelling test (and later, arithmetic and language tests) to large numbers of pupils in selected

[49]Fass, op. cit., footnote 25, p. 433. See also Cremin, op. cit., footnote 2, p. 100.

[50]See, e.g., Gould, op. cit., footnote 43, for a fuller discussion of the role of testing in the eugenics movement and how it influenced public policy in the 1920s and 1930s.

[51]Fass, op. cit., footnote 25, p. 432.

cities. His findings, published in 1892, were based on data he had collected on some 30,000 children, and documented the absence of a relationship between the time schools spent on spelling drills and children's performance on objective tests of spelling.[52] ' 'In one study, [Rice] . . . found that [instructional time] varied from 15 to 30 minutes per day at different grade levels . . . [but that] tests of student performance on a common list of words revealed that the extra 15 minutes a day made no difference in demonstrated spelling ability."[53] When Rice's results were presented to a major meeting of school superintendents in 1897, they were ridiculed; ultimately, however, a few farsighted educators concurred with Rice's analysis.[54]

Managerial Efficiency

Schools were not alone in their attempts to adapt to changing times. The following description of change in the railroad industry could just as well describe emerging trends in school administration:

> . . . it meant the employment of a set of managers to supervise . . . functional activities over an extensive geographical area; and the appointment of an administrative command of middle and top executives to monitor, evaluate, and coordinate the work of managers responsible for the day-to-day operations. It meant, too, the formulation of brand new types of internal administrative procedures and accounting and statistical controls. . . .[55]

In other sectors of American enterprise, engineers, researchers, and managers were applying scientific principles to enhance efficiency. In agriculture, for example, research and technology was transforming the nature and scale of farming. Progressive educators, who were familiar with the commercial precedents, ''. . . commonly used the increased productivity of scientific farming as an analogy for the scientifically designed educational system they hoped to build. ' '[56]

The newly evolving business organizations also employed modes of classification and bureaucratic control that bore remarkable similarity to those adopted by school systems as they shifted from largely rural, decentralized organizations to urban, centralized ones. "Scientific management, a relatively late addition to the set of new business organizational principles invented around the turn of the century, was based on the proposition that managers could ascertain the abilities of their workers and assign them accordingly to the jobs where they would be the most productive.

Managerial efficiency was but one way in which business thinking coincided with school policy. The other principal point of convergence had to do with the demand for "skilled' labor. Just as division of labor according to ability was seen as a vehicle to improve productivity on the shop floor, classification and ranking of students was seen as a prerequisite to their efficient instruction. The relationship is perhaps best illustrated by the statements of Harvard President Charles Eliot, in 1908. Society, he said, is:

> . . . divided. . . into layers. . . [with] distinct characteristics and distinct educational needs . . . a thin upper [layer] which consists of the managing, leading, guiding class . . . next, the skilled workers . . . third, the commercial class . . . and finally the thick fundamental layer engaged in household work, agriculture, mining, quarrying, and forest work. . . . [The schools could be]. . . reorganized to serve each class. . . to give each layer its own appropriate form of schooling.[57]

It was an obvious leap, then, for business executives to join with progressives in calling for reform of schools along the corporate model. Hierarchy, bureaucracy, and classification—all served by the science of testing-would become the institutional environment charged with producing educated persons capable of functioning in the hierarchical, bureaucratic, and classified world of business.[58]

---

[52] Haney, op. Cit., footnote 3, p. 600.

[53] Resnick, op. cit., footnote 3, p. 180.

[54] Monroe, op. cit., footnote 44, pp. 88-89.

[55] Alfred Chandler, *The Visible Hand: The Managerial Revolution in American Business* (Cambridge, MA: Harvard University Press, 1977), p. 87. Chandler's description of changes in railroad school administration. Daily reports-from conductors, agents, and engineers-detailed every aspect of railroad operations; these reports, along with information from managers and department heads, were used to make day-to-day decisions and, at the executive level, to compare the performance of operating units with each other and with other railroads (p. 103).

[56] Tyack and HanSot, Op. Cit., footnote 9, p. 157.

[57] Tyack, op. cit., footnote 4, p. 129.

[58] For a critical analysis of testing and social/economic stratification in the United States, see, e.g., Clarence Karier, ''Testing for Order and Control in the Corporate Liberal State," in Block and Dworkin (eds.), op. cit., footnote 24, pp. 339-373.

The advocates of the corporate model of school governance, such as Stanford Education Dean Ellwood P. Cubberley, argued that to manage efficiently, the modem school superintendent needed "rich and accurate flows of information' on enrollments, buildings, costs, student promotions, and student achievement.[59] Cubberley advocated the creation of ''scientific standards of measurement and units of accomplishment" that could be applied across systems and used to make comparisons. Fulfilling this need for data, Cubberley maintained, would require new types of school employees—efficiency experts ". . . to study methods of procedure and to measure and test the output of its works";[60] a recommendation that indeed came to pass as large, urban systems hired census takers, business managers, and eventually evaluation experts and psychologists.

Achievement and Ability Vie for Acceptability

Despite initial opposition from teachers, the use of achievement tests as instruments of accountability began to gain support. By 1914 the National Education Association was endorsing the kind of standardized testing that Rice had been urging for two decades. The timing was exquisite: on one front, there was the "push" of new technology that promised to be valuable to testing, and on the other, a heightened "pull" for methods to bring order to the chaotic schools.

Two approaches to testing competed for dominance in the schools in the early 20th century. One had its antecedents in the intelligence testing movement, the other in the more curriculum-oriented achievement testing that grew out of Rice's examples.

Between 1908 and 1916, Edward Thorndike and his students at Columbia University developed standardized achievement tests in arithmetic, handwriting, spelling, drawing, reading, and language ability. Composed of exercises to be done by students, the arithmetic test was similar in format to the types of tests traditionally administered by teachers. The handwriting and composition tests, by contrast, consisted of samples of handwriting and essays against which pupil performances were compared.[61] By 1918, there were well over 100 standardized tests, developed by different researchers to measure achievement in the principal elementary and secondary school subjects.[62]

Student achievement was not all that would come under the microscope of standardized assessment. In the frost decade of the 20th century, following the advice of Cubberley and other advocates of scientific management, ". . . leaders of the school survey movement examined and quantified virtually every aspect of education, from teaching and salaries to the quality of school buildings."[63] Indeed, Thorndike's proclamation of 1918--"whatever exists at all exists in some amount'—formed the cornerstone of his educational measurement edifice.[64] By 1922, John Dewey would lament the victory of the testers and quantifiers with these words: "Our mechanical, industrialized civilization is concerned with averages, with percents. The mental habit which reflects this social scene subordinates education and social arrangements based on averaged gross inferiorities and superiorities."[65]

Thorndike's approach to achievement tests mirrored in important ways that taken by reformers in Massachusetts some 70 years earlier: just as they had reached a foregone conclusion about the quality of Boston schools before the frost tests were given, Thorndike's tests actually came ***after*** he had already decided that the schools were failing. His 1908 study of dropouts, followed the next year by a remarkable statistical analysis conducted by Leonard Ayres,

---

[59] Tyack and Hansot, op. cit., footnote 9, p. 157.

[60] Ellwood P. Cubberly, *Administration* (Cambridge, MA: The Riverside Press, 1916), p. 338.

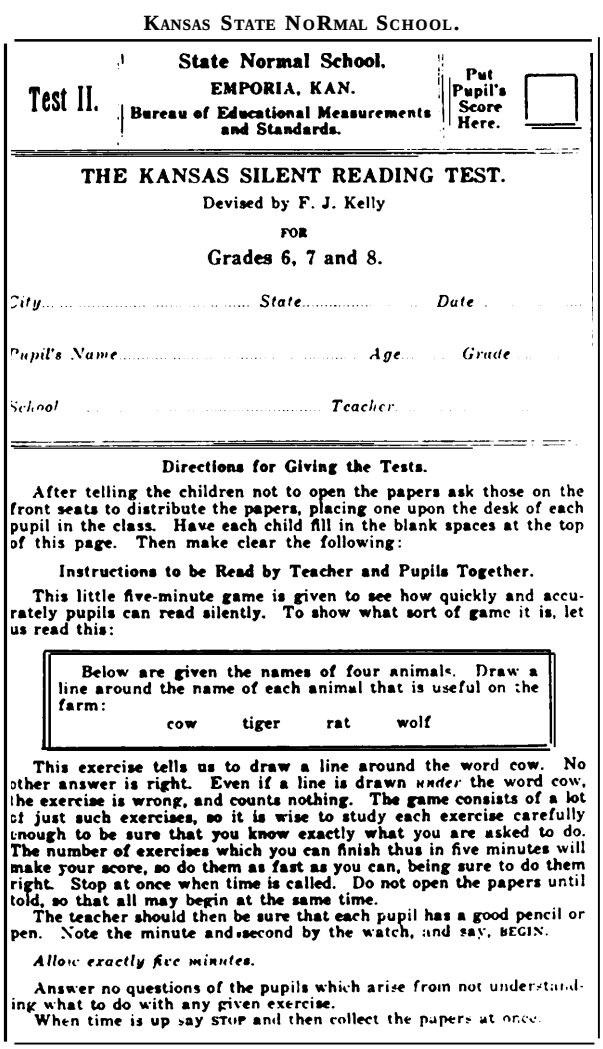
[61] Monroe, op. cit., footnote 44, p. 90.

[62] Cremin, op. cit., footnote 2, p. 187, A report by Walter Monroe in 1917 documented over 200 such tests. See Chapman, oP. cit., footnote 29, p. 34.

[63] Chapman, op. cit., footnote 29, pp. ,35,,

[64] In later writings, Thorndike was more humble. For example, he wrote: ''Existing instruments (for measuring intellect) represent enormous improvements over what was available twenty years ago, but three fundamental defects remain. Just what they measure is not known; how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained is not known; just what the measures obtained signify concerning intellect is not known. . . .' Edward L. Thorndike, E.O. Bregman, M.V. Cobb, and Ella Woodyard, York, NY: Columbia University, Teachers College, Bureau of Publications, 1927).

[65] Tyack, op. cit. footnote 4, p. 198.



KANSAS STATE NORMAL SCHOOL.

State Normal School.
EMPORIA, KAN.
Bureau of Educational Measurements and Standards.

Test II.

Put Pupil's Score Here.

THE KANSAS SILENT READING TEST.

Devised by F. J. Kelly

FOR

Grades 6, 7 and 8.

City..... State..... Date .....

Pupil's Name..... Age..... Grade .....

School..... Teacher.....

Directions for Giving the Tests.

After telling the children not to open the papers ask those on the front seats to distribute the papers, placing one upon the desk of each pupil in the class. Have each child fill in the blank spaces at the top of this page. Then make clear the following:

Instructions to be Read by Teacher and Pupils Together.

This little five-minute game is given to see how quickly and accurately pupils can read silently. To show what sort of game it is, let us read this:

Below are given the names of four animals. Draw a line around the name of each animal that is useful on the farm:

cow tiger rat wolf

This exercise tells us to draw a line around the word cow. No other answer is right. Even if a line is drawn under the word cow, the exercise is wrong, and counts nothing. The game consists of a lot of just such exercises, so it is wise to study each exercise carefully enough to be sure that you know exactly what you are asked to do. The number of exercises which you can finish thus in five minutes will make your score, so do them as fast as you can, being sure to do them right. Stop at once when time is called. Do not open the papers until told, so that all may begin at the same time.

The teacher should then be sure that each pupil has a good pencil or pen. Note the minute and second by the watch, and say, BEGIN.

Allow exactly five minutes.

Answer no questions of the pupils which arise from not understanding what to do with any given exercise.

When time is up say STOP and then collect the papers at once.

Photo credit: Rutgers University Press

The first educational test using the multiple-choice format was developed by Frederick J. Kelly in 1915. Since then, multiple choice has become the dominant format of standardized achievement tests.

called attention to an alarming problem.[66] For reasons that neither Thorndike nor Ayres professed to understand entirely, the schools were full of students who were not progressing. In New York City, for example, Ayres reported that 23 percent of the 20,000 children studied were above the normal age for their grade.

Where could concerned educators of the time turn for explanations? It is useful to review in this context the staggering demographic changes of the time, a phenomenon that so utterly consumed the collective psyche that Thorndike, Ayres, or anyone else thinking about the schools could not have helped but try to explain their findings in terms of the changing national origin of students. Between 1890 and 1917, the total U.S. population grew from 63 million to over 100 million, largely as a result of immigration. During the same period, the population aged 5 to 14 grew from just under 17 million to over 21 million; similarly, the public school enrollment rate climbed from about 50 percent in 1900 to 64 percent in 1920, and average daily attendance went from 8 million to just under 15 million.[67]

The effects of immigration and population growth on the issues Thorndike and Ayres grappled with, however, were somewhat surprising. While Ayres's initial research question— "Is the immigrant a blessing or a curse?"[68]--reveals something about the anti-immigrant zeitgeist, his answers, based on the data analysis he presented, revealed a healthy objectivity. Ayres concluded that:

1. there was no evidence that the problems of students being above normal age for their grade or dropping out were most serious in those cities having the largest foreign populations;
2. "... children of foreign parentage drop out of the highest grades and the high school faster than do American children;
3. ... there are more illiterates among the native whites of native parentage than among the native whites of foreign parentage; "[69] and
4. "... the proportion of children five to fourteen years of age attending school is greater among

[66] See Leonard Ayres, *Laggards in Our Schools: A Study of Retardation in City Systems* (New York, NY: Russell Sage Foundation, Charities Publication Committee, 1909), p. 8.

[67] For analysis of the effects of child labor laws on school attendance, see David Goldston, History Department, University of Pennsylvania, "To Discipline and Teach: Compulsory Education Enforcement in New York City, 1874-94," unpublished monograph, n.d.

[68] Ayres, op. cit., footnote 66, p. 103.

[69] Ibid., p. 115. Ayres did not cite the source for his illiteracy statistics, which he presumably collected himself. Census data suggest a somewhat different picture from the one presented by Ayres. In 1900, for example, about 5 percent of the native white population was estimated to be illiterate, as compared to almost 13 percent of the foreign born. Had Ayres included the census category "Negro" (and other races), he might have found-as did the census-a staggering illiteracy rate of 44 percent in 1900. See Bureau of the Census, op. cit., footnote 5, Series H 664-668, p. 382.

those of foreign parentage and foreign birth than among Americans.' '[70]

" Finally, he concluded from his analysis that: " . . . in the country at large [the schools] reach the child of the foreigner more generally than they do the child of the native born American,' which was a source of great humiliation to ''national pride. ' '[71]

Experimentation and Practice

Although Ayres may not have been aware of it, his work actually vindicated the basic tenets of the achievement-oriented testers, who tended to focus on school curricula and the extent to which children were actually mastering the substantive content of schooling. Their approach to assessment was to develop quantitative and qualitative measures of student ''productions;' and the ''. . . early versions of standardized tests were developed by public school systems, often in collaboration with university centers, to reflect the curriculum of the schools in a particular city."[72]

This approach to assessment recognized implicitly that institutional factors were largely responsible for the sorry situation in the schools. Moreover, if school practices changed, then children's opportunities for success would improve, and it was believed that the kind of information provided by the standardized achievement tests could light the way to effective reform.

Much to the frustration of the dedicated educators who had mounted them, the effects of school reform efforts were typically disappointing. In New York, for example, in 1922, nearly one-half of all students were above the normal age for their school grade, and there was enormous variability in ages of pupils in any given grade.[73]

This sort of experience did not dissuade educators from the idea of using tests to effect change, but rather persuaded many of them that poor student achievement stemmed from low innate ability. In other words, even the achievement tests of Thorndike were inadequate to measure-and remedy—the problems of schools, because those tests did not adequately measure basic intelligence. The statements of New York Superintendent William Ettinger underscore the intrinsic appeal of the intelligence test model:

> . . . rapid advance in the technique of measuring mental means that we stand on the threshold of a new era in which we will increasingly group our pupils on the basis of both intelligence and accomplishment quotients and of necessity, provide differentiated curricula, varied modes of instruction, and flexible promotion to meet the crying needs of our children.[74]

Thus, for Ettinger and others, the achievement tests available at the time were still not standardized enough—they did not get at the root causes of difference in student performance.

New York was not alone. Oakland, California, was the site of one of the first attempts at large-scale intelligence testing of students. During the 1917 and 1918 academic years, 6,500 children were given the Stanford-Binet, as well as a new test written by Arthur Otis (one of Lewis Terman's students who would eventually be credited with the invention of the multiple-choice format[75]). The experiment in Oakland was significant because it was one of the frost attempts to use intelligence tests to classify students: "Intelligence tests were used at frost to diagnose students for special classes; later their adoption led to the creation of a systemwide tracking plan based on ability. . . . The experiment with testing in Oakland . . . would provide a blueprint for the intelligence testing movement after the war.' '[76]

The Influence of Colleges

Another institutional force exerted pressure on the schools during this period. The university sector sent a clear message of dissatisfaction with the quality of high school graduates, and urged a return to the high standards to which the elite colleges had been accustomed in earlier times. Many academic leaders

---

[70]Ayres, op. cit., footnote 66, p.115.

[71]Ibid., p. 105.

[72]Edward Haertel and Robert Calfee, "School Achievement: Thinking About What to Test,'' vol. summer 1983, p. 120.

[73]Tyack, *op.* cit., footnote p. 203.

[74]Ibid.

[75]See ch. 8.

[76]Chapman, op. cit., footnote 29, p. 56.

were attracted to the intelligence test as a filter in their admissions process. The President of Colgate, along with leaders of the Carnegie Foundation, the University of Michigan, Princeton, Lehigh, and other higher education institutions, argued that too many children were in college who did not belong there.

As early as 1890, Harvard President Charles William Eliot proposed a cooperative system of common entrance examinations that would be acceptable to colleges and professional schools throughout the country, in lieu of the separate examinations given by each school. The interest of Eliot and like-minded college presidents in a standardized set of national examinations went beyond their immediate admissions needs. Their broader objective was to institute a consistent standard that could be used to gauge not only the quality of high school students' preparation, but also, by inference, the quality of the high schools from which those students came. The ultimate aim was to prod public secondary schools to standardize and raise the level of their instruction, so that students would be better prepared for higher education. Eliot expressed consternation that. . . in the present condition of secondary education one-half of the most capable children in the country, at a modest estimate, have no open road to colleges and universities. "[77]

Getting colleges and universities to agree on the subjects to be included and the content knowledge to be assessed in a common college entrance examination was no easy task. Anticipating the minimum competency testing movement by almost a century, the opponents of a standard college entrance examination voiced early concerns about whether these tests could lead to State examinations that would eventually be used for awarding degrees as well as college admission.

Eventually the advocates of common examinations were able to garner enough support to form the College Entrance Examination Board in 1900. In 1901, the first examinations were administered around the country in nine subjects. While in later years college admissions examinations would come to resemble tests of general intelligence, the early examinations of the College Board were closely tied to specific curricular requirements: ''. . . the hallmark [of the examinations] was their relation to a carefully prescribed area of content. . . .' '[78]

Within a relatively short period of time, the College Board became a major force on secondary school curricula. The Board adopted the practice of formulating and publicizing, at least a year before a new examination was introduced, a statement describing the preparation expected of candidates. Developed in consultation with scholarly associations, these statements, in the opinion of one observer, ''. . . became a paramount factor in the evolution of secondary school curriculum, with a salutary influence on both subject matter and teaching methods. ' '[79] This glowing assessment was not shared by all educators. By the end of World War I, many school superintendents shared the concerns of one California teacher who wrote the following to the Board in 1922:

> These examinations now actually dominate, control, and color the entire policy and practice of the classroom; they prescribe and define subject and treatment; they dictate selection and emphasis. Further, they have come, rightly or wrongly, to be at once the despot and headsman professionally of the teacher. Slight chance for continued professional service has that teacher who fails to ''get results' in the "College Boards, " valuable and inspiring as his instruction may otherwise be.[80]

World War I

Army testing during World War I ignited the most rapid expansion of the school testing movement. In 1917, Terman and a group of colleagues were recruited by the American Psychological Association to help the Army develop group intelligence tests and a group intelligence scale. This later became the Alpha scale, used by the Army to quickly and efficiently determine which recruits were capable for service and to assign them to jobs.[81]

---

[77] Much of the discussion of the early history of the College Board comes from John A. Valentine, The *College Curriculum: Influence Standards* York, NY: College Entrance Examination Board, 1987). Eliot is quoted on p. 3.

[78] From the autobiography of James B. Conant, quoted in ibid., p. 21.

[79] Claude M. Fuess, quoted in ibid., p. 19.

'Ibid., p. 29.

81 Monroe, op. cit., footnote 44, p. 95.

The administration of group intelligence tests during the war stands out to this day as one of the largest social experiments in American history. Prior to World War I, most intelligence tests had been administered to individuals, not large groups. In a period of less than a month, the Army's psychologists developed and field tested an intelligence test. Almost as quickly, the Army began applying the tests to what today would clearly be called "high-stakes decisions. ' The Alpha tests, for the normal population, and the Beta tests, for the subnormal, both loosely structured after Binet's tests for children, were given to just under 2 million young Army men, and the results were used as the basis for job assignments. "In short, the tests had *consequences: in* part on the basis of a short group examination created by a few psychologists in about a month, testee number 964,221 might go to the trenches in France while number 1,072,538 might go to offices in Washington. ' '[82]

The results from this testing were mixed. For one thing, validation studies were less than conclusive and Army personnel (and others) criticized the validity of the tests. In one such study (the typical validation study used officers' ratings of soldiers' proficiencies as the outcome or criterion measure), correlations between performance on the Alpha test and officers' ratings were in the low 0.60s, and on the Beta test in the 0.50s.[83] The Army itself had mixed feelings about the testing program, and eventually it discontinued testing its peacetime force.

One of the most important outputs of the program was the mass of data that could be mined by eager intelligence theorists. Some theorists reached particularly controversial and inflammatory conclusions, most notably that 1) a substantial proportion of American soldiers were "morons," which was presented as evidence that the American "stock" was deteriorating; and 2) in terms of test performance, the ranking of intelligence was white Americans first, followed by Northern Europeans in second place, with immigrants from Southern and Eastern Europe a distant third. These findings helped fuel the work of a small but vocal group of eugenicists, such as Carl Brigham, who advocated

> . . . selective breeding [to create] a world in which all men will equal the top ten percent of present men. . . "[84]

This reasoning contributed to congressional debate over restrictive immigration legislation.[85]

Testing Through World War II: 1918 to 1945

Overview

Several themes emerged during the period of 1918 to 1945 that continue to be relevant to testing policy. A basic lesson of the period was that in a society constantly struggling with tradeoffs between equity and efficiency, an institution that claims to serve both objectives at once commands attention. If achievement and intelligence tests had been viewed purely in terms of more efficient classification, they would have undoubtedly encountered even more public opposition than they did. But because the tests were promoted as tools to aid in the efficient allocation of resources according to principles of

---

[82] Tyack, op. cit., footnote 4, p. 204.

[83] A 0.5 validity coefficient does not m-that predictions of soldiers' future performance based on their test scores were right about one-half the time. Rather, it suggests a linear and nonrandom relationship (O conflation would signify complete randomness) between the score and the criterion variable. It should be noted that today's tests used for selection and placement (e.g., the Scholastic Aptitude Test for college admissions or the General Aptitude Test Battery for employment) have predictive validities (correlation coefficients) in the 0.2 to 0.4 range. See, e.g., Frank Hartigan and Alexandra Wigdor, *Employment* National Academy Press, 1989). For a critique of the policy to use employment tests with low predictive validity, see, e.g., Henry Levin, "Issues of Agreement and Contention in Employment Testing, " December 1988, pp. 398-403.

Since the days of the Army Alpha, the psychometric quality of tests used in screening and selection has improved considerably; in fact, there is little evidence that the criterion measures for the Army Alpha were psychometrically sound, or that other test features would pass today's scientific muster. Stephen Jay Gould made this point quite forcefully in his book *Mismeasure of Man* (op. cit., footnote experiment that demonstrated how Harvard students, hardly an illiterate lot, performed on the Beta version of the test-designed for recruits who could not read-is often cited as prima facie evidence of the low psychometric quality of the Army intelligence tests.

[84] Karier, op. cit., footnote 58, p. 347. Some of the early faith in eugenics was fueled by the writing of H.H. Goddard, as described in Gould (op. cit., footnote Fancher (op. cit., footnote 43), and other histories. However, it is important to note that Goddard later recanted his findings concerning the allegedly low intelligence levels of immigrants and Black Americans, and publicly apologized for the effects those findings might have had. For discussion see Carl Degler, *In of Nature* (Cambridge, England: Oxford University Press, 1991).

[85] See, e.g., Gould, op. cit., footnote 43.

''meritocracy, ' they appealed to a wide spectrum of the American polity.[86]

Second, the development of mental measurement—part of the broader emergence of psychology as a bona fide science--coincided with profound demographic and geographic shifts in American society. New educational testing models were cultivated in this crossroads of technological push (psychology) and social pull (the need to reform schools and schooling). Windows of opportunity of this sort are rare in history; how society capitalizes on them can have deep and long lasting impacts.

Third, it is important to distinguish technology of testing from ideology of test use. The history of testing in America suggests that political, social, and economic uses for testing can substantially exceed the technical limits imposed by test design.[87]

Fourth, there appears to be a trend from highly specific and curriculum-oriented achievement tests toward tests of increasingly general cognitive ability. This trend has historically been associated with attempts to extend principles of accountability to larger and larger jurisdictions, i.e., from schools to districts to States and ultimately to the Nation as a whole. As shown by the developments in college admissions testing, for example, the move toward consolidation of admissions criteria and the perceived need to influence secondary school education nationwide led eventually to the adoption of a test designed explicitly to assess aptitude, which later was renamed "developed ability," rather than achievement of specific curricular goals. This trend has been reinforced, historically, by several other factors:

. the incentives for efficiency, made particularly important by the commitment to assess massive numbers of students over many different learning objectives;
- the recurring interest in using tests as a way to mitigate the cultural differences in a heterogeneous population; and
- the tendency to shift blame for the quality of education, i.e., to explain low achievement in terms of low innate ability of students rather than in terms of poor management and instruction.

Fifth, growth in the use of standardized tests often coincides with heightened demand for greater unification in curricula. Although the history does not demonstrate a fixed direction of causality, it does suggest the following sequence: initially there is growing recognition that many schools are not doing as well as they should; next there is awareness of a fragmented school system which, if nothing else, makes it difficult to obtain systematic information about what is really happening in classrooms; and finally there is a simultaneous push for standardization in measurement-to facilitate reliable comparisons and standardization of instruction--to remedy the fragmentation.

A Legacy of the Great War

Despite the questionable foundations and effects of the Army's intelligence testing experiments, the terrain had been plowed, and on the conclusion of World War I, schools were only too willing to partake of the harvest. At long last, it seemed to many school leaders, there was a technology that could be deployed in the service of elevating the quality of education provided to the Nation's youth. ''Better testing would allow [the schools] to perform their sifting scientifically,"[88] i.e., to classify children according to their innate abilities and in so doing, protect the slow witted from the embarrassments of failure while allowing the gifted to rise to their rightful levels of achievement.

World War I, in effect, set in motion the process that would result-in an incredibly short time-in

---

[86] The word ''meritocracy'' was coined the English sociologist Michael Young in satirical essay. See Young, *of the Meritocracy, 1870-2033: on* (London, England: Thames and Hudson, 1958). Paula Fass notes that: "The IQ established a meritocratic standard which seemed to sever ability from the confusions of a changing time and an increasingly diverse population provided a means for the individual to continue to earn his place in society by his personal qualities, and answered the needs of a sorely strained school system to educate the mass white locating social talent.' Fass, op. cit., footnote 25, p. 446.

[87] Historian Michael Katz disagrees:

> I can't agree with. . . the point . . . that there's a difference between the purpose of testing (or the technology or science of testing) and the uses to which testing is put. . . . This argument creates a false dichotomy which seems to reflect a naive view of scientific and technological development as self-contained and unaffected by their context. Clearly, this wasn't so; psychology and testing as research enterprises were products of time and place with all that implies.

Katz, op. cit., footnote 42.

[88] Tyack, op. cit., footnote 4, p. 206.

national intelligence testing for American school children. By the end of the first decade after the war, standardized educational testing was becoming a fixture in the schools. A key development of the period was the publication of test *batteries,* which

> . . . relieve[d] the teacher or other user from the task of selecting the particular tests to be used . . . [and which provided] a method for combining the several achievement scores into a single measure. " Many testmakers included detailed instructions and scoring procedures for using achievement and intelligence tests in conjunction with each other, in order to gauge". . . how well a school pupil is capitalizing his mental ability."[89]

The proponents of testing were extraordinarily successful: ''. . . one of the truly remarkable aspects of the early history of IQ testing was the rapidity of its adoption in American schools nationwide."[90] Another aspect was that researchers obtained their data not from a controlled laboratory or limited trial programs, but from real schools in which millions of students were taking the tests. This period of testing, then, involved a complicated two-way interaction between the research community and the public, with the mass testing of children-and the use of test results to support important administrative decisions--occurring even as research on the validity and usefulness of tests continued to develop.

It is not surprising that testing engendered public controversy, given that its most visible manifestation in those days was in *selection.* Had the tests been used to diagnose learning disorders among children and to **create** appropriate interventions, they would have likely enjoyed more public support. But the tests were mostly used as they had been during the war, namely to classify (i.e., label and rank) individuals, and to assign them to positions accordingly. A U.S. Bureau of Education Survey conducted in 1925 showed that intelligence and achievement tests were increasingly used to classify students.[91] Group-administered intelligence tests were most likely to be used for classification of pupils into homogeneous groups, and educational achievement tests were most likely to be used to supplement teachers' estimates of pupils' ability. Related survey data showed that 90 percent of elementary schools and 65 percent of high schools in large cities grouped students by ability, and that the use of intelligence tests as the basis for classification was widespread.

By the fall of 1920 the *World Book* had published nearly half a million tests, and by 1930 Terman's intelligence and achievement tests (the latter published as the Stanford Achievement Test) had combined sales of some 2 million copies per year. If test production and sales are any indicator of social preferences, the data suggest a marked preference for achievement measures over tests of innate intelligence. Between 1900 and 1932, there were some 1,300 achievement tests on the market, as compared to about 400 tests of "mental capacities. ' '[92] High school tests, vocational tests, assessments of athletic ability, and a variety of miscellaneous tests had been developed to supplement the intelligence tests, and statewide testing programs were becoming more common.[93]

The Iowa Program

In 1929, the University of Iowa initiated the first major statewide testing program for high school students. Directed by E.F. Lindquist, the Iowa program had several remarkable features: every school in the State could participate on a voluntary basis; every pupil in participating schools was tested in key subjects; new editions of the achievement tests were published annually; and procedures for administering and scoring tests were highly structured. Results were used to evaluate both students and schools, and schools with the highest composite achievement received awards. In addition, Lindquist was among the first to extend the range of student abilities tested. The Iowa Tests of Basic Skills and the Iowa Test of Educational Development became tools for diagnosis and guidance in grades three to eight and in high school, respectively. The Iowa program was also a significant demonstration of the feasibility of wide-scale testing at a reasonable cost.

89 Monroe, op. cit., footnote 44, p. 99.

90 Fass, op. cit., footnote 25, p. 445.

91 W.S. Deffenbaugh, Bureau of Education, U.S. Department of the Interior, "Uses of Intelligence Tests in 215 Cities," City School Leaflet No. 20, 1925.

92 Chapman, op. cit., footnote 29, (citing data from Hildreth), p. 149.

93 Monroe, op. cit., footnote 44, pp. %, 106, and 111.

E.F. Lindquist (1901-1978), at left, one of the fathers of standardized achievement testing, directed the Iowa testing programs. In 1952, E.F. Lindquist developed the basic circuitry design for the first electronic scoring machine, as shown below.

Photo credit: University of Iowa Press

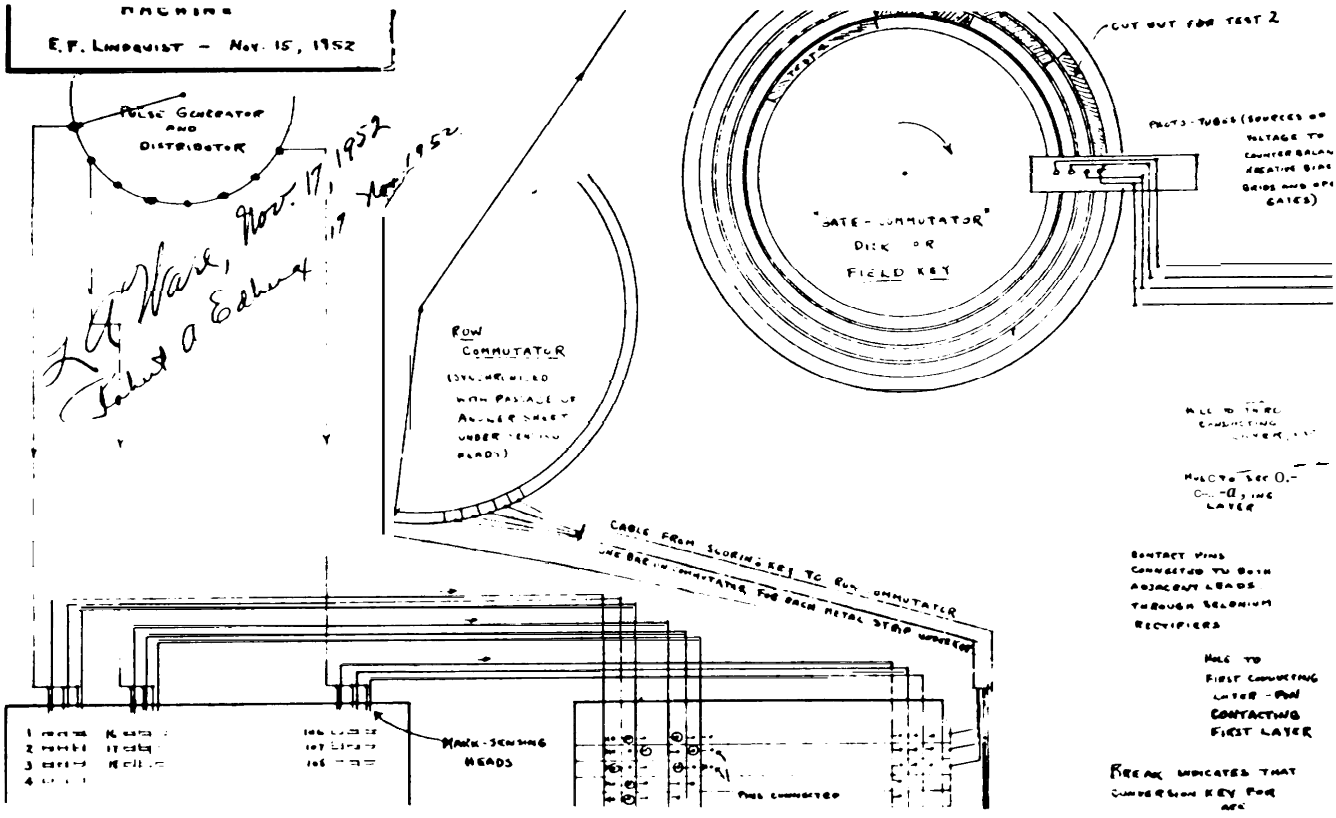


Photo credit: University of Iowa Press

By the late 1930s, Iowa tests were being made available to schools outside the State.[94]

Under Lindquist, the Iowa program had a remarkable influence on swinging the pendulum of educational testing back in the direction of diagnosis and monitoring, and away from classification and selection. Indeed, the distinction between intelligence and standardized achievement tests, in their design and content as well as their scores, was always fuzzy. In any event, the use of intelligence tests encountered substantially heavier criticism than the use of achievement tests—if not on the grounds of their relative design strengths and weaknesses, then on the extent to which they became the basis for classifying and labeling children early in their lives.

Multiple Choice: Dawn of an Era

The achievement tests that gained popularity during the 1920s looked very different from the pre-World War I educational tests. Achievement tests were designed largely with the purpose of sorting and ranking students on various scales. This model of test design has dominated achievement testing ever since.

One of the most significant developments was the invention of the multiple-choice question and its variants. The Army tests marked the first significant use of the multiple-choice format, which was developed by Arthur S. Otis, a member of the Army testing team who later became test editor for *World Book.* In the view of the Army test developers, the multiple-choice format provided:

> . . . a way to transform the testees' answers from highly variable, often idiosyncratic, and always time-consuming oral or written responses into easily marked choices among fixed alternatives, quickly scorable by clerical workers with the aid of superimposed stencils.[95]

The multiple-choice item and its variant, the true-false question, were quickly adapted to student tests and disseminated for classroom use, marking another revolution in testing. Lindquist and coworkers at the Iowa program later invented mechanical and later electromechanical scoring machines that would make possible the streamlined achievement testing of millions of students.[96]

Not surprisingly, the rapid spread of multiple-choice tests kindled debate about their drawbacks. Critics accused them of encouraging memorization and guessing, of representing "reactionary ideals" of instruction, but to no avail. Efficiency and ''objectivity' won out; by 1930 multiple-choice tests were firmly entrenched in the schools.

Critical Questions

In the late 19th and early 20th centuries, the potential for science to liberate the schools from their shackles of inefficiency was almost universally accepted. As suggested earlier, this fact helps explain the apparently ironic marriage of testing and progressivism.

But if the spirit of progressivism catapulted scientific-style testing, it was that same progressivism that ultimately reined it in. In a nutshell, the intelligence testers went too far. When Brigham used the Army data to argue that Blacks were naturally inferior; when Robert Yerkes wrote that one-half of the white recruits were morons; when H. H. Goddard suggested that the intellectually slovenly masses were about to take over the affairs of state; or when a popular writer named Albert Wiggam " . . declared that efforts to improve standards of living and education are folly because they allow weak elements in the genetic pool to survive, [and] that 'men are born equal' is a great 'sentimental nebulosity' . . ";[97] it became clear to progressives like John Dewey that testing had run amok.

Thus, in the days immediately following the first World War, the "heyday of intelligence testing" was confronted by a kind of field day of antitesting muckraking. And the muckrakers were progressives: most notably, Walter Lippman, whose 10 articles in the New *Republic* attempted to remind readers that

> . . . the Army Alpha had been designed as an

---

[94] Julia J. Peterson, *Iowa* (Iowa City, IA: University of Iowa Press, 1983), pp. 1-6.

[95] Franz Samelson, ''Was Early Mental Testing (a) Racist Inspired, (b) Objective Science, (c) A Technology for Democracy, (d) The Origin of Multiple Choice Exams, (e) None of the Above? Mark the RIGHT Answer,' *Society:* M. Sokal (cd.) (New Brunswick, NJ: Rutgers University Press, 1987), p. 116.

[96] See discussion in ch. 8.

[97] Cronbach, op. cit., footnote 40, p. 9. For a Comprehensive survey of the questionable scientific basis for intelligence testing, see Gould, op. cit., footnote 43.

instrument to aid classification, not to measure intelligence. '[98] It was almost as though Lippman, an early supporter of tests to aid in the efficient management of schools, suddenly recognized that the very same tests could be put to different ends. "Intelligence testing," Lippman warned, "could . . . lead to an intellectual caste system in which the task of education had given way to the doctrine of predestination and infant damnation."[99]

College Admissions Standards: Pressure Mounts

The admissions procedures established by the College Board had some clearly beneficial effects on education. They succeeded in enforcing some degree of uniformity in the college admissions process, helped raised the level of secondary school instruction, engendered serious discussion about the appropriate curriculum for college-bound youth, and built solid, cooperative relationships among higher education institutions throughout the country.[100]

Nevertheless, several influential colleges continued to express concern that most secondary schools did not take the mission of college preparation seriously and did not organize their curricula within the College Board's guidelines. Moreover, despite the board's energetic efforts at standardization, a large portion of the Nation's colleges continued to rely to some extent on their own examinations.[101]

In addition, college leaders were coming to a more sophisticated recognition of the limitations of achievement-type tests, including the College Board tests, in helping admissions officers discriminate between students who had stockpiled memorized knowledge and students with more general intellectual ability. Harvard was particularly sensitive to the apparently high number of applicants who, ". . . as a result of constant and systematic cramming for examinations . . . manage to gain admission without having developed any considerable degree of intellectual power."[102] Partly in response to this problem Harvard developed a plan that in a fundamental way presaged the eventual swing from curriculum-centered achievement tests toward more generalized tests of intellectual ability: the plan called for a shift from separate subject examinations to "comprehensive" examinations designed to measure the ability to synthesize and creatively interpret factual knowledge.

At Columbia University, as well, the pressure was on to do something about the admissions process. The arrival of increasing numbers of immigrants, many of them Eastern European Jews living in New York City, fueled the xenophobia. Columbia's President, Nicholas Butler, for example, found the quality of the incoming students (in 1917) ". . . depressing in the extreme . . . largely made up of foreign born and children of those but recently arrived. . . ."[103] To counteract this trend, Butler adopted the Thorndike Tests for Mental Alertness, hoping that ". . . would limit the number of Jewish students without a formal policy of restriction."[104]

In 1916, the College Board began developing comprehensive examinations in six subjects. These examinations included performance types of assessment such as essay questions, sight translation of foreign languages, and written compositions. While the comprehensive examinations enabled colleges to widen the range of applicants, university leaders continued to watch with interest the development and growing acceptance of intelligence tests.

Responding to the demand for standardization and for tests that could sort out applicants qualified for college-level work from those less qualified, the College Board developed the Scholastic Aptitude Test (SAT). The test was administered for the first time in 1926; one-third of the candidates who sat for College Board examinations took the new test, and the SAT was off to a promising start.[105]

---

[98] Cremin, op. tit., footnote 2, p. 190.

[99] Walter Lippman, "The Abuse of the Tests,'' p.

[100] Valentine, op. cit., footnote 77, p. 17.

[101] Ibid., pp. 32-33.

[102] Claude M. Fuess, The *College Board: Its First Fifty Years* (New York, NY: College Entrance Examination Board, 1%7), quoted in ibid., p. 24.

[103] Harold Wechsler, *The Qualified Student* (New York, NY: John Wiley, 1977), *p.* 155.

[104] James Crouse and Dale Trusheim, The *Case Against the SAT* (Chicago, IL: University of Chicago Press, 1988), p. 20. See also Resnick, Op. Cit., footnote 3, p. 188.

[105] Valentine, op. cit., footnote 77, p. 35.

In addition to reinforcing the growing popularity of multiple-choice items, the SAT made several other contributions to the testing enterprise. First, the College Board took pains to try to prevent misinterpretation of SAT results. The board's manual for admissions officers cautioned that the new tests could not predict the subsequent performance of students with certainty and further warned of the pitfalls of placing too much emphasis on scores. Second, the board also adopted procedures from the outset to ensure confidentiality of test scores and examination content.[106] Third, the unique scoring scale, from 200 to 800, with 500 representing the average, indicated where students stood relative to others, a concept that helped lay the underpinnings for the eventual dominance of norm-referenced testing.

Given the central role of colleges and universities in American life generally and their specific influence on secondary education standards, it is perhaps not surprising that examinations designed for selection soon became the basis for rather general judgments about individuals' ability and achievement, or that in later years, the SAT would become the basis even for inter-State comparisons of school systems. Clearly the SAT was not designed or validated for either of those purposes,[107] as its designers have attempted to clarify time and again; the fact that it was appropriated to those ends, therefore, stands out as a warning of how tests can be misused.

Testing and Survey Research

Along with the increased use of standardized tests for tracking in the elementary and secondary grades and for college admissions, the period between the wars also saw the first uses of standardized tests in large-scale school surveys. These studies, which paved the way for the kinds of program evaluations that would become so important in education policy analysis in the 1960s, had several aims. Researchers, journalists, and charitable foundations seized on surveys as a way of calling attention to inequities and shortcomings in public education. Understandably, these studies met resistance from school superintendents, who resented being called on the carpet by outsiders. But as the old guard of superintendents were gradually replaced by people more familiar with the role of quantitative analysis in educational reform, and as superintendents came to see the benefits of an outside inventory of school needs, particularly in terms of increased public support for more funding, attitudes softened.[108]

The links between achievement test scores and later college performance were further challenged by Ralph Tyler's analysis of data generated in the "Eight-Year Study" (1932 to 1940).[109] In looking for evidence of a link between formal college-preparatory work in high school and eventual college performance, Tyler reached several important conclusions. First, his research revealed that certain basic tenets of the progressive movement, e.g., deemphasizing rigid college entrance requirements in the high school curriculum, did not produce graduates who were less well prepared for college work than those in traditional classrooms. Second, Tyler's research ". . . confirmed the importance of following student progress on a continuous basis, recording data from standardized tests as well as other kinds of achievement. ''[110] Third, it set an important precedent for the use of achievement scores as a control variable in large-scale survey-based studies. Finally, the study demonstrated the

---

[106] Ibid., pp. 31-37.

[107] The Scholastic Aptitude Test is intended as a source of additional information, over and above high school grades, to predict freshman grade point average. While its predictive validity has been documented, even that rather modest mission-as compared with overall judgments of individual ability or State education systems—is controversial. See, for example, Crouse and Trusheim, op. cit., footnote 104.

[108] Tyack and Hansot, op. cit., footnote 9, p. 163.

[109] The study involved a group of 30 public and private secondary schools, which had been invited to revise substantially their course offerings and provide a more flexible learning environment for students intending to go to college. Cooperating with these 30 schools were some 300 colleges and universities that had agreed to waive their formal admissions requirements. Tyler examined the effects of high schoolwork on college performance among 1,475 pairs of students—each consisting of a graduate of one of the 30 schools and a graduate of another school not in the study, matched as closely as possible on race, sex, age, aptitude test scores, and background variables.

[110] Resnick, op. cit., footnote 3, page 186.

potential power of educational research as an agent of change.[111]

Another development in the years between the wars was high-speed computing, first applied to testing in 1935. Although there was by then little argument with the idea of standardized testing, the cost-effectiveness of using electronic data processing equipment to process massive numbers of tests was icing on the cake. One report showed that the cost of administering the Strong Inventory of Vocational Interests dropped from $5 per test to $.50 per test as a result of the computer.[112]

Testing and World War II

Once again, new research ground was broken on the eve of world war. But unlike the experience with the Army Alpha program in World War I, the testing that took place during the second World War did not substantially affect educational testing; nor did it engender much public controversy. For one thing, testing was already so well ensconced in the public mind-several million standardized tests were administered annually by the outbreak of the war-that the testing of 10 million Army recruits hardly seemed out of the ordinary. Second, the Army testing program did not focus on innate ability and the hereditarian issue. And third, it did not seem to rest on assumptions of a unitary dimension of intelligence. Rather, it seems that the theoretical and empirical studies initiated by Thurstone, Lindquist, and others had succeeded in persuading the Army psychologists to consider alternative models with which to estimate soldiers' abilities and future performance.

"Multiple assessment," which examined distinct mental abilities, such as verbal comprehension, word fluency, number facility, spatial visualization, associative memory, perceptual speed, and reasoning, was one of two significant technological developments in testing during this period.[113] Another was the transfer of testing technology from the schools to the military. For example, elements of the Iowa Tests of Basic Skills and the Iowa Test of Educational Development were borrowed by the Army for their World War II testing program, establishing the credibility of tests based on notions of multiple dimensions of ability.

Equality, Fairness, and Technological Competitiveness: 1945 to 1969

Overview

Much of the controversy over student testing during the post-World War II period revolved around its uses in classification and selection. Although there had always been some dissent, controversy over student testing had entered a relatively quiet phase in the late 1920s, allowing the psychometric community to refine its craft and the educational community to create ". . . the most tested generation of youngsters in history. ' '[114] But astute listeners in the early post-war years could detect faint rumblings of conflict; by the end of the 1960s testing would once again be in the eye of storm over educational and social policy.

Three sets of forces came to bear on the schools in general and on testing policy in particular during the 1950s and 1960s: demographic change, due largely to new immigration, which once again challenged the American ideal of progressive education; technological change, brought into sharp relief by the launching of Sputnik, which ignited nation-

---

111 Commenting on the Eight-Year Study, Lee Cronbach and Patrick Suppes wrote:

> Although the study was carried out as planned, one cannot escape the impression that the central question was of minor interest to the investigators and the educational community. The main contribution of the study was to encourage the experimental schools to explore new teaching and counseling procedures.

Lee Cronbach and Patrick Suppes (eds.), *Research for* York, NY: MacMillan Publishing Co., 1%9), pp. 66-67. George Madaus (personal communication, 1991) notes that the Eight-Year Study was a turning point in the design of tests: it supported Tyler's argument that direct measures of performance needed to precede the design of indirect measures. See also G. Madaus and D. Stufflebeam (eds.), *Educational Evaluation: Classical Works Tyler* (Boston, MA: Kluwer, 1989).

112 Resnick, op. cit., footnote 3, p. 190. For more discussion of the technology of testing see ch. 8.

113 To this day, the debate between the unitary and multidimensional intelligence theorists rests in stalemate, largely because each camp uses different mathematical models to analyze test scores. As Howard Gardner has neatly pointed out: "Given the same set of da@ it is possible, using one set of factor-analytic procedures, to come up with a picture that supports the idea of a 'g' factor; using another equally valid method of statistical analysis it is possible to support the notion of a family of relatively discrete mental abilities. ' Howard Gardner, 2nd ed. (New York, NY: Basic Books, 1985), p. 17, and ch. 6 of this report.

114 Cremin, op. cit., footnote 2, p. 192. Daniel and Lauren Resnick would later embellish this theme, arguing that ' 'American children were the most tested in the world-and the least examined.'' See Daniel P. Resnick and Lauren Resnick, "Standards, Curriculum and Performance: A Historical Perspective," vol. 14, No. 4, April 1985, p. 17.

Photo credit: Marjory Collins

Testing of children has often involved oral as well as written work. These first grade pupils at the Lincoln School of Teachers' College, Columbia University, are recording their voices for diction correction, circa 1942.

wide interest in science and mathematics education as well as higher standards of schooling overall; and the awakening of the public conscience to the problems of racial inequality in the Nation's public schools, which led to wholly new approaches to school governance, financing, and participation.

Access Expands

Enrollment in public elementary and secondary schools jumped from 25 million in 1949-50 to 46 million in 1969-70, or from 17 percent of the total population to over 22 percent. The number of high school graduates went from just over 1 million in 1950 to 2.6 million in 1970. The trend was even more impressive in the postsecondary sector: total enrollments in institutions of higher education went from 2.6 million in 1949-50 to 8 million in 1969-70. While part of the enrollment growth is explained by the size of the "baby boom" cohort, the increase in the proportion of the population enrolled in school signifies progress toward the goal of universal access.

The timing of this upsurge in participation suggests that through decades of increased reliance on standardized tests, the progressive spirit in American education had not only survived, but had actually flourished. Several points need to made in this regard. First, recall that student classification had been viewed by the early progressives as a means to render schooling more efficient: it was when tests became designed and used to classify students on the basis of innate ability-and to allocate educational resources accordingly-that some of the Progressives began to protest. Although the proponents of testing could argue that their approach was intended to ensure continued high standards of school quality, the resulting sorting and tracking of children was anathema to many leaders of the Progressive movement (Dewey, in particular).[115]

Second, both sides claimed to have the welfare of children and the Nation at heart. It was commonly agreed that schooling needed to improve; the dispute arose over the choice of strategy. One side favored increased access to education by all students, and tolerated or supported testing as a way to manage massive public education more efficiently. The implicit assumption was an egalitarian one: all children could learn. The other side also favored testing; but the underlying assumption was that some children were innately more capable of learning than others, and that classification would keep standards high for the more able students while

[115] On the acceptability of testing by the Progressive movement, see also Cronbach, op. cit., footnote 40, p. 8. While Cronbach concedes that the testers themselves may have gone too far in their reliance on the new science of measurement, he seems to place more of the blame for controversy on the popular press: "Virtually everyone favored testing in schools; the controversies arose because of incautious interpretations made by the testers@ even more, by popular writers."

sparing the slower ones the embarrassment of failure.[116]

The Test of General Educational Development (GED) played an interesting role in expanding educational access. The GED was formulated by the U.S. Armed Forces Institute, in cooperation with the American Council on Education, to address the problems of returning service personnel who had been inducted before graduating from high school. Patterned after the Iowa Test of Educational Development and constructed with substantial input from Lindquist, the GED was intended to enable out-of-school youth and adults to demonstrate knowledge for which they would receive academic credit and in some cases a high school equivalency diploma.[117]

Thus, the postwar enrollment boom and the development of the GED could be viewed as a victory for universal access. But the analysis would be remiss without repeating the obvious: these developments took place in an education culture fully infused with standardized tests. Indeed, it would be possible to argue-as some did-that tests *opened gates of opportunity,* that access to school was enhanced, not encumbered, by objective tests.[118] In later years this theme would be echoed by some minority leaders, who argued that standardized tests allowed children the opportunity to demonstrate their ability more effectively—and more fairly—than they had been able to in the highly subjective environments of their impoverished classrooms.[119] This curious nature of testing-it could be assigned responsibility for enhancing or for confining opportunities for advancement-sheds light on its powerfully symbolic role in American society generally and in education specifically.

Developments in Technology

American enchantment with technology during the 1950s produced several strides in the field of testing. Most noteworthy was the automatic scoring machine, a form of optical scanner invented by the Iowa Testing Program. The machine enabled tests to be processed in large volume and at a reasonable cost.[120] During the next 12 years, the Iowa program, through its engineering spinoff, the Measurement Research Center, perfected several generations of scanners, each smaller but more powerful than the last.[121]

With this equipment, national testing programs became feasible. Although the optical scanning equipment did not in itself drive up demand for testing, it gave an efficiency edge to tests that could be scored by machine and enabled school systems to implement testing programs on a scale that had previously been unthinkable. An enormous jump in testing ensued. One estimate of the number of commercially published tests administered in 1961

---

[116] The tension between access and standards has been a longstanding motif in education policy debates. Lawrence Cremin illustrates it eloquently in his summary of former Harvard President James Conant's conflicted views on the subject:

> For Conant . . . the mixing of youngsters from different social backgrounds with different vocational goals in comprehensive high schools is important to the continued cohesiveness and classlessness of American society, important enough to maintain in the face of the difficulty of providing a worthy education to the academically talented in the context of that mixing. Hence, the central problem for American education is how to preserve the quality of the education of the academically talented in comprehensive high schools.

Cremin, op. cit., footnote 2, p. 23.

[117] Peterson, Op. cit., footnote 94, p. 82.

[118] Christopher Jencks and David Reisman argue that the 'Conservatives ''in the debate over college admissions policies were those who disliked tests and who preferred the old-fashioned criteria (e.g., that sons of alumni should be granted preference); and that the "liberals" were those who favored ''. . . seeking out the ablest students. . . wherever they might come horn.' They goon to suggest that while the liberals appear to have been winning, there has been a". . . rising crescendo of protest, especially from the civil rights movement and others who believe in a more egalitarian society, against the use of tests to select students and allocate academic resources. " See their seminal work, *The Academic Revolution* (New York, NY: Doubleday, 1%9) pp. 121 ff.

[119] See, e.g., Donald Stewart, ''Thinking the Unthinkable: Standardized Testing and the Future of American Education,'' speech before the Columbus Metropolitan Club, Columbus, OH, Feb. 22, 1989. Stewart, who is president of the College Entrance Examination Board, notes that:

> In a country as multicentric and pluralistic as ours, only a standardized test that works like the SAT is going to be valuable. . . in providing some . . . national sense of the levels of educational ability of different individuals and also different groups.

He goes on to note that:

> . . . the SAT has made it possible for students from every background and geographic origin to attend even the most prestigious institutions.

[120] Peterson, op. cit., footnote 94, p. 89.

[121] Ibid., p. 163.

was 100 million[122]--just under 3 tests per year, on average, for each student enrolled in grades K- 12.[123]

In 1958, Iowa also introduced computerization to the scoring of tests and production of reports to schools. This early and rather primitive application of computers to the field of testing helped propel two decades of research and development that culminated in highly sophisticated programs of computer-based testing.

But technology played an important role not just in the design and implementation of tests, but as a catalyst to renewed interest in the use of testing to improve education. By the mid-1950s, a major expansion in educational opportunities was taking place amid a continued reliance on standardized tests to diagnose and classify students and monitor school quality. The impetus for this expansion came in large part from America's rude awakening to global technological advance: the Soviet launching of Sputnik (Oct. 4, 1957) spurred many Americans to question whether the battlefield victories in World War II were sufficient for America to win the peace that followed. As in prior periods of perceived external challenge, the policy response centered on education, and as in prior periods, the education reforms involved increased testing. The general idea behind the National Defense Education Act of 1958 was to provide Federal funds for upgrading mathematics and science education in particular.

One means for accomplishing this goal was the allocation of Federal dollars to support the development and maintenance of:

> . . . a program for testing aptitudes and abilities of students in public secondary schools, and . . . to identify students with outstanding aptitudes and abilities . . . to provide such information about the aptitudes and abilities of secondary school students as may be needed by secondary school guidance personnel in carrying out their duties; and to provide information to other educational institutions relative to the educational potential of students seeking admissions to such institutions. . . [124]

Race and Educational Opportunity

The birth of the modem civil rights movement was a watershed in American history and marked a turning point in the history of schooling. It also altered the course of testing policy and raised new debates about the design and use of various tests in school and the workplace.

In 1954, the *Brown v. Board of Education* Supreme Court decision ruled out racial segregation in schools, thereby establishing the legal prescription for completing the mission of the public school movement. It had taken about 100 years to address this glaring anomaly in a school system predicated on the ideal of universal access. *Brown* had no immediate and direct consequences for testing, but it set in motion social and ideological forces that would, in years to come, bring student testing into new arenas of controversy and, for the first time, into the courts.

In a second significant court case, *Hobson v. Hansen (1967),* filed on behalf of a group of Black students in Washington, DC, the policy of using tests to assign students to tracks was challenged on the grounds that it was racially biased. The judge concurred; although the test was given to all students, the court found that because the test was standardized to a white, middle class group, it was inappropriate to use for tracking decisions.[125]

The explicit rejection of the notion of "separate but equal' in *Brown* set the tone for challenges such as *Hobson,* which found that tests used for classification could result in the kinds of racially segregated classrooms (or schools) explicitly outlawed by *Brown.* A new branch of applied statistics emerged, concerned with the analysis of group differences in test scores in order to determine the potential ''adverse impact' of test use in certain kinds of decisions.

---

[122]David Goslin, The *Search for Ability* York, NY: Russell Sage, 1963).

[123]Total K-12 enrollments in the *1959.60* school year were just over 36 million. See U.S. Department of Education, Digest *Education Statistics,* 1990 (Washington, DC: U.S. Government Printing Office, 1991), p. 47.

[124]National Defense Education Act, Public Law 85-864.

[125]269 F. Supp. 401 (D.D.C. 1967).

Controversies emerged over the effects of tests in correcting or exacerbating racial inequality.[126] Two other points need to be made about this period. First, the civil rights movement led to the development of a wide range of social programs, which in turn created new demands for accountability measures to ensure that Federal money was being well spent. A century after accountability became a purpose of student testing at the State and local level, the model was being applied on a grand scale to national issues. The 1965 Elementary and Secondary Education Act in particular opened the way for new and increased uses of norm-referenced tests to evaluate programs.

Second, controversy over the quite obvious increased reliance on testing for selection and monitoring decisions did not abate; on the contrary, even the notion of using certain kinds of ability tests to classify children into categories such as ''educably mentally retarded,' for the purpose of giving them special educational treatment, came under strident criticism by parents and leaders who viewed the classification as potentially harmful to their children's long-term opportunities.

Recapitulation

Testing of students in the United States is now 150 years old. From its earliest incarnation coinciding with the birth of mass popular schooling, testing has played a pivotal role in the American experiment with democratic education. That experiment has been unique in many ways. Not only did it begin well before most other industrialized countries expanded schooling to the masses, but it was carried out in a uniquely American, decentralized system: today 40 million children attend schools scattered across some 15,000 local school districts. If there have been taboos in American education, they have concerned national curriculum, national standards, and national testing.[127]

Yet for all its diversity, the American system also shows some remarkable uniformity and stability. Beneath the surface of institutional independence lies a strong unifying force, a tacit agreement that a principal objective of schooling is *community:* "E pluribus unum" does not stop at the schoolhouse door. But neither does it come with a handy recipe to make it work. Indeed, the apparently endless struggle over the structure, content, and quality of American education—and of educational tests—stems in part from the tension between the judgments of teachers, parents, and students on the one hand, and the quest for community, State, or even national standards, on the other.

Teachers in their classrooms have always used all kinds of tests-everything from spot quizzes to group projects—as part of the continuous process of assessment of individual student learning. At the same time, as this chapter has shown, standardized examinations have been used at least since the mid- 19th century to keep district and State education authorities, and the legislatures that fund them, informed about the general quality of schools and schooling. From their inception, these tests have been used to inform institutional decisions about student placement and resource allocation, and they have been seen as a way to influence teaching and learning standards.

Today the United States stands again at the crossroads of major transition in student testing. The issues framing today's public policy debate--perceived decline in academic standards, shifts in the demographic composition of the student population, heightened awareness of global technological competition, and lingering inequality in the allocation of educational and economic opportunities—have been evolving for two centuries. Lessons from the history of educational testing provide important background to the development of testing policies for the future.

---

[126] The most vehement debate was sparked by the 1969 publication of an article by Arthur Jensen questioning whether school intervention programs (such as Head Start) could affect IQ, which was largely determined by heredity. See Arthur Jensen, ''How Much Can We Boost IQ and Achievement?' *Harvard Educational* vol. 39, winter 1969. For review of this controversy see, e.g., Cronbach, op. cit., footnote 40; Mark Snyderman and Stanley Rothman, *IQ* Brunswick, NJ: Transaction Books, 1988); and Fancher, op. cit., footnote 43.

[127] This picture is changing. See discussion in ChS. 1 and 2 of this report.