This article was downloaded by: [171.66.48.206]
On: 23 April 2013, At: 14:25
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of Curriculum Studies
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tcus20

Retrospective on educational testing and assessment in the 20th century

Marguerite M. Clarke , George F. Madaus , Catherine L. Horn & Miguel A. Ramos

Version of record first published: 08 Nov 2010.

To cite this article: Marguerite M. Clarke , George F. Madaus , Catherine L. Horn & Miguel A. Ramos (2000): Retrospective on educational testing and assessment in the 20th century, Journal of Curriculum Studies, 32:2, 159-181

To link to this article: http://dx.doi.org/10.1080/002202700182691

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

J. CURRICULUM STUDIES, 2000, VOL. 32, NO. 2, 159–181

Retrospective on educational testing and assessment in the 20th century

MARGUERITE M. CLARKE, GEORGE F. MADAUS, CATHERINE L. HORN and MIGUEL A. RAMOS

Over the last 100 years, the ever-increasing demand for testing as a measure of educational reform has created a very profitable market for the US testing industry. We follow the growth of this market since the 1900s in two different, but related, ways. First, we discuss some of the technical developments that have encouraged the use of standardized testing in general and contributed to the growth of the commercial testing industry. Second, we attempt to quantify the expansion of the testing marketplace during the 20th century by tracking several indirect indices of growth over time. We conclude that although technical innovations may have contributed to the growth of the US testing marketplace, they do not necessarily lead to better tests or better outcomes for those who use them. There is a need to more carefully monitor the effects of these tests on teaching and learning in general, particularly when the tests are used in high-stakes contexts.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Since the turn of the century, standardized commercial tests have been widely used to measure the achievement of US students (Madaus *et al.* in press). Over the decades, despite vigorous criticism from some quarters, these tests have been widely regarded as administratively convenient, inexpensive tools that could help solve an array of educational problems (National Commission on Testing and Public Policy 1990). In fact, most educational reforms now rely heavily on testing to serve a multitude of purposes: to show increased rigour of school curricula; to determine if students advance or graduate; to judge the effectiveness of schools and teachers; and to compare districts, states and nations (Bennett 1999, National Research Council [NRC] 1999).

Calculating just how many standardized tests US students currently sit for is a complex job, involving such questions as what defines a 'test',

Marguerite M. Clarke, Campion Hall, Boston College, Chestnut Hill, MA 02467, USA, is a research associate with the National Board on Educational Testing and Public Policy and a doctoral candidate in educational research, measurement and evaluation at Boston College.
George F. Madaus, a senior fellow with the National Board on Educational Testing and Public Policy, is the Boisi Professor of Education and Public Policy at Boston College. He is the former director of Boston College's Center for the Study of Testing, Evaluation and Public Policy.
Catherine L. Horn is the current Boisi Graduate Fellow and a doctoral candidate in educational research, measurement and evaluation at Boston College.
Miguel A. Ramos is a doctoral candidate in educational research, measurement and evaluation at Boston College.

Journal of Curriculum Studies ISSN 0022–0272 print/ISSN 1366–5839 online © 2000 Taylor & Francis Ltd
http://www.tandf.co.uk/journals/tf/00220272.html

whether a test battery is counted as one test or several, and which student population figures to use. One estimate, which included state-mandated testing programmes, local district testing programmes, tests for special populations (e.g. special education), and college admissions testing, is that elementary and secondary students take between 140 million and 400 million tests per year (Haney *et al.* 1993). Given that these figures are over five years old, and given the steady increase in state-mandated testing programmes, current estimates would probably be closer to the upper end of this range.

The demand for testing as a measure of educational reform combined with the ever-increasing numbers of students taking these tests makes for a highly profitable market (Hoffman 1962). We follow the growth of this market for testing over the last century in two different, but related, ways.[1] First, we examine some of the technical developments that have contributed to the increased use of standardized testing in the US and the social context in which they took place.[2] We look at the effect of these developments on what standardized tests have measured over time and how they have measured it. We also describe the way in which these developments have contributed to the growth of the commercial testing industry itself. In the second part of the paper, we attempt to quantify the extent of this expansion of the testing marketplace during the 20th century by tracking several indirect indices of growth over time. These indices include references to testing in the *Education Index* (H. W. Wilson 1932–1998), the dollar amount of standardized test sales, and the revenues of the major testing companies. The discussion throughout is limited to commercial, standardized achievement tests produced for the US elementary and secondary market—otherwise known as the 'Elhi' market. This market also includes tests for college admissions—such as those produced by the College Entrance Examination Board (College Board) and American College Testing (ACT)—because these are primarily administered to students in high school.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Technical developments

Whatever noun is chosen—assessment, examination, or just plain test—they all encompass the same basic concepts and the same basic technology (Madaus 1993, 1994, 1995). Thus, while our title uses the terms 'testing' and 'assessment', we view these terms as interchangeable. 'Test' was the word of choice for much of the 20th century. 'Assessment', however, is the favoured term of the 1990s, either when used alone or when modified by one of the adjectives 'authentic', 'alternative' or 'standards-based'.

Although the enterprise of testing is itself a technology, here we focus on specific technical developments in the areas of test development, format and scoring that have contributed to the rise of standardized testing in the US since the 1900s.[3] Developments such as the invention of the multiple-choice format in 1914, the optical scanner in the 1950s and the rise of computer-adaptive testing (CAT) in the 1990s have impacted how the domain to be tested is operationalized and how test scores are interpreted.

Two points are worth making here. First, for the most part, developments in the technical hardware of testing have been part of a move toward increased efficiency that has characterized standardized achievement testing for much of this century. This move led to the proliferation of multiple-choice standardized achievement tests from the 1920s on and made feasible the federal and state legislation of large-scale district and statewide multiple-choice testing programmes over the last half of the century (Madaus 1994). In the late 1980s, the rise of the alternative assessment movement with its preference for open-ended questions and performance assessments seemed to call into question this historical preoccupation with efficiency and standardization. However, the efficiency imperative is still at work, as witnessed by the large numbers of states that use multiple-choice and short-answer items as their predominant assessment modes—albeit supplemented by open-ended and performance items (Council of Chief State School Officers [CCSSO] 1998). In addition, some testing programmes that initially relied heavily on open-ended and performance items (e.g. the Kentucky Instructional Results Information System) have come full circle and now rely heavily, if not entirely, on multiple-choice and short-answer items (Kentucky Department of Education 1996, Reidy 1997). Whether testing will fully revert to its efficiency imperative is as yet uncertain. Either way, there will be definite repercussions for the commercial industry that has grown up around testing and uses the efficient multiple-choice item as its format of choice. Second, the rise of computer technology has had a definite impact on the enterprise of testing in general and on the commercial industry in particular. The introduction of computers into almost every aspect of testing—from item and test development to administration to scoring and reporting—has the potential to influence what tests measure and how they measure it. In addition, specific developments such as computer-based tests (CBT) and CAT have great revenue potential if certain costly aspects such as item development and security issues can be addressed. Thus, the future of testing holds great promise, but also great uncertainties for many of those in the industry. We will return to these issues later. We focus here on technical developments in the testing enterprise in the US over the last century and the social contexts in which they took place. Our discussion is divided into two periods: pre- and post-World War II.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Pre-World War II

The early decades of the 20th century were characterized by a common faith in the power of technology, quantification, a benign science, a culture of objectivity, and cool reason to solve all manner of social problems (Postman 1992, Porter 1995, Kanigel 1997). Two test-related developments in particular—the introduction of the IQ test and the invention of the multiple-choice format—were products of the spirit of the time and provided the impetus for the growth in standardized testing that began during this period.

Although we do not deal here directly with intelligence testing, it is useful to point out its impact on the rise of the standardized testing industry. During a good deal of the 19th century, when there were relatively small numbers of students, there was a belief that they all could learn if properly taught (Horace Mann Papers, 1845–1846, United Kingdom 1886). This belief began to erode in the face of the poor performance of a larger and much more diverse population of students. The advent of the IQ test—designed by Binet for French school children—in the USA just prior to World War I permitted educators to shift blame for poor attainment away from teaching toward a lack of students' 'ability to profit from instruction'. For example, in 1918 Charles Hubbard Judd, then Director of the School of Education at the University of Chicago, contended that 'unsatisfactory school results [are] to be traced to the native limitations in the ability of the child or to the home atmosphere in which the child grows up' (Judd 1918: 152). He (1918: 153) went on to extol the virtues of 'scientific' measurement:

> We all understand now in definite scientific terms that children are different from one another . . . . that the best we can hope for is improvement—not absolute achievement of ideals. With the theoretical ideal of perfection overthrown, there is now an opportunity to set up rational demands. We can venture to tell parents with assurance that their children in the fifth grade are as good as the average if they misspell fifty percent of a certain list of words. We know this just as well as we know that a certain automobile engine cannot draw a ton of weight up a certain hill. No one has a right to make unscientific demand of the automobile or of the school.

'Scientific' tests of both achievement and intelligence quickly began to serve as selection devices to identify talent and to place students in the 'proper' curriculum for their ability level.[4] However, Binet's design for the IQ test had serious bureaucratic drawbacks. His scales had to be individually administered, scored and interpreted by trained psychologists. Simply put, the technology was inefficient; it did not lend itself to widespread use for screening and grouping.

With the advent of World War I, the US government needed a way of efficiently classifying recruits. It turned to a group intelligence test developed by Otis that became known as the *Army Alpha* and *Beta* (one for literates, one for illiterates) and by 1918, these tests were being administered to 2 million recruits. Haney (1984) notes that this need to classify recruits did for psychological testing what microwaves did for the processed-food industry 60 years later. By 1932, 75% of 150 large city school systems in the USA used group intelligence tests to track students into ability groups; colleges also used tests to rationalize admissions procedures, and the results were often used for exclusionary purposes (Haney 1984).

The key to this large-scale efficient testing programme was the use of the multiple-choice item invented in 1914 by Frederick J. Kelly (Samelson 1987). Many more items could now be administered in a short period, and tests could be scored quickly and objectively by unskilled clerks. The multiple-choice design was widely adopted by the fledgling test-publishing industry that emerged in the 1920s and evolved into a billion dollar

Downloaded by [171.66.48.206] at 14:25 23 April 2013

enterprise. (The highly profitable nature of this industry will be discussed in greater detail later). By 1922, the boast was made that 'most of the tests [on] the market, unless measuring handwriting, do not call for written answers' (Pressey and Pressey 1922: 186).

Ralph Tyler's work on behavioural objectives in the 1930s and 1940s further impacted the fledgling test industry. Tyler insisted that while early test-developers described their test domains in terms of content to be covered, they overlooked the types of cognitive responses that should be expected of pupils. He maintained that because a pupil could perform the indicated operations upon a given list of numerical exercises or could solve a word problem, did not necessarily mean the test was valid. Tyler became the champion of defining educational objectives in terms of both a content and a behavioural component (Tyler 1934). His idea that educational objectives should contain a behavioural component eventually led to the development of Bloom *et al.*'s (1956) *Taxonomy of Educational Objectives: The Cognitive Domain*.[5] Tyler's ideas were to also provide a foundation for the work of Mager (1962) and Popham (1978) who championed behavioural objectives testing in the 1960s and 1970s.

Several of the major players in the US Elhi (i.e. elementary and high school) market were founded during these early decades—e.g. the College Board in 1900, Houghton Mifflin in 1916, and the Psychological Corporation (PsychCorp), California Test Bureau and World Book in the 1920s. Taken together, the invention of the multiple-choice format which made it possible to produce tests in volume, the societal drive towards efficiency and scientific management which created a market for these tests, and the increased popularity of intelligence testing which supplied a justification for the way these tests were interpreted and used, provided a conducive environment for the fledgling industry. For example, the need for some uniform standard that would aid in making college admissions decisions more efficiently gave rise to the founding of the College Board in 1900. The Board's first common college entrance tests, in essay form, were administered beginning in 1901. After the widespread publicity given to intelligence testing with multiple-choice examinations in the early 1920s (Haney 1984), the Board appointed a committee of experts to advise it on the suitability of developing multiple-choice tests for use in college admissions. This resulted in the administration of the new Scholastic Aptitude Test (SAT), for the most part multiple-choice in format, to 8000 candidates in June, 1926. By the late 1940s, the essay examinations were completely abandoned, partly because of the appearance of studies showing that the multiple-choice SAT could predict college performance as well as essay examinations.

World Book, one of the earliest and most successful test publishers, entered the market in the 1920s with the production of achievement and intelligence tests by Arthur Otis and Lewis Terman. By 1930, yearly sales of the Otis/Terman group intelligence test by World Book were over US $750 000 and those of the Stanford Achievement Tests (Stanford) were US$1.5 m. (Chapman 1980: 111–112).[6] Another publisher who entered the test-publishing business with the production of intelligence tests was Houghton Mifflin, who began publication of the Stanford-Binet Intelli-

Downloaded by [171.66.48.206] at 14:25 23 April 2013

gence Scale in 1916. Since then, Houghton Mifflin has concentrated test publishing in its Riverside Press subsidiary which now produces two of the most popular achievement tests in the US Elhi market—the *Iowa Tests of Basic Skills* and the *Iowa Tests of Educational Development.*

Post-World War II

Beginning in the late 1950s, four social forces combined to create an expanding market for standardized testing in the US (Haney *et al.* 1993). First was recurring public dissatisfaction with the quality of education, and several concomitant waves of educational reform. Witness the Sputnik uproar of the 1950s, followed by the 'basic skills' movement of the 1970s, the release of the National Commission on Excellence in Education's *A Nation at Risk* in 1983, and finally, the *Goals 2000: Educate America Act* in the 1990s. In each of these 'reform' waves, testing was seen as an important policy tool. Second was an array of federal and state legislation promoting or explicitly mandating standardized testing programmes, beginning with the National Defense Education Act of 1958. Third was a broad shift in attention, signalled by the famous Coleman report (Coleman *et al.* 1966), from evaluating the inputs or resources devoted to education to measuring the outputs or results—operationalized by student performance on available multiple-choice tests. Finally, increased bureaucratization of US society in general, and of schooling in particular (Wise 1979), made the technology of multiple-choice, standardized, commercial tests an attractive tool. Tests provided a means for categorizing people, educational institutions, and problems according to abstract, impersonal and generalizable rules and helped expedite formal and impersonal administrative procedures. These four factors were intimately related one to the other; for example, public dissatisfaction with the quality of US education produced legislation that in turn contributed to increased bureaucratization (Haney *et al.* 1993). Although each of the last five decades of the 20th century can be defined in terms of the operation of one or more of these influences, the 1990s have been a time when all four came into play, creating an unprecedented level of testing activity in the US.

Apart from the effects of these social forces on the demand for tests in the US, several technical innovations combined to spur the use of standardized testing from the 1950s on. Perhaps one of the most significant innovations was the invention of the high-speed scanner in 1955. This invention, coupled with the already popular multiple-choice format, led to increased efficiency and reduced the cost of testing. It then became possible to test every student in a state for about two or three dollars per student, and have the results back in two to four weeks. During the 1960s and 1970s, multiple-choice testing proliferated in the form of state-mandated minimum competency testing. Figure 1 shows the growth in numbers of US states authorizing minimum-competency tests and assessment programmes from 1950 through 1985. During this time period, there was a steady rise in numbers of state-mandated assessment programmes—from one in 1960 to 35 by 1985. The rise in numbers of states authorizing minimum compe-

Downloaded by [171.66.48.206] at 14:25 23 April 2013

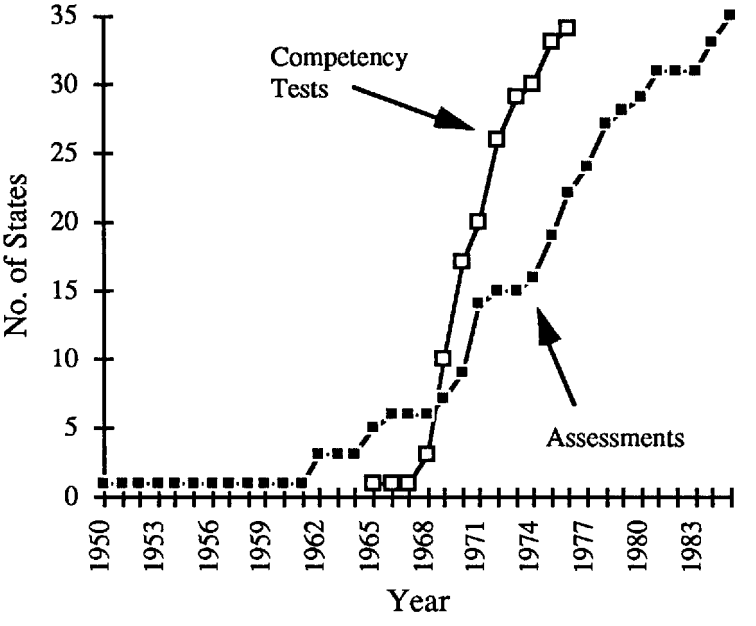

Figure 1. Numbers of states authorizing assessment programmes and minimum competency testing, 1950–1985 (Sources: Office of Technology Assessment 1987, Haney *et al.* 1993).

tency tests was even more dramatic. The curve rises sharply from one such programme in 1967 to 34 by 1976. Naturally, with every state mandate, the number of students tested—and hence the number of tests administered—increased. By 1985, state testing programmes were the most important market for publishers' group achievement and ability tests (Fremer 1989).

Several testing companies were to capitalize on the invention of the high-speed scanner. For example, National Computer Systems (NCS), which was founded as a corporation in Minnesota in 1962 and became a publicly held company in 1968, dealt only in test and survey scoring services from the outset. In fact, according to Holmen and Docter (1972), test-scoring initially accounted for 70% of NCS's data-processing activity, and scoring of survey forms for the remaining 30% Since then, NCS's role has expanded to include test-publishing and contracted test-development. However, NCS continues to derive its revenues not so much from its own tests as from the testing work it does for others—scoring tests, selling test-scoring machines, and doing testing contract work for state agencies and others. Scantron is another example of a company that capitalized on the scanner. Scantron was founded 10 years after NCS and in the mid-1970s introduced the first desktop scanner to read and score test answer sheets via microcomputer (Baker 1989). (The company went public in 1983 and became a subsidiary of John Harland Co. in 1988). Scantron now has the largest market share of optical mark-reading school-building-level equipment. However, it makes its profit primarily from the scannable forms that schools must purchase when they are given a rent-free machine.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Developments in techniques for gathering and analyzing information on large numbers of examinees (e.g. matrix sampling and item response theory) also led to a rapid increase in the number of large-scale assessment or testing programmes from the 1970s on. These large-scale assessments were conducted at the state, national and international levels. From the late 1980s on, two developments in particular had an effect on the nature of these large-scale assessments at the state level: the rise of the alternative assessment movement, and the increased tendency for US states to mandate academic content standards and assessments aligned with these standards. The next section addresses these two developments in greater depth, including their impact on the commercial test industry.

The move towards content standards and assessments aligned with those standards is an important sea-change in testing and has implications for the testing industry in terms of the adaptations they must make to changing market needs. Part of the impetus for content standards came from the announcement by former-President Bush and the National Governors' Association of six National Education Goals to be achieved by the year 2000. Linked to these goals was a call for new 'American Achievement Tests' covering 'core subjects' like English, mathematics, and science, and based on 'new world standards' (US Department of Education 1991). Curriculum groups were subsequently formed in different subject areas to develop national content standards on which such tests might ultimately be based. A number of states also began to develop their own standards in different subject areas and to align their assessments with these standards. This produced a new potential market for test publishers because states often did not have the expertise or technical hardware in-house to assume this kind of project. Since the early 1990s, the number of states that have assessments aligned with their standards has increased dramatically, providing a lucrative market for test publishers. For example, Texas has paid NCS about US$20 million per year for developing its test—the Texas Assessment of Academic Skills. NCS in turn subcontracts this test to Harcourt Brace Educational Measurement. Texas also pays Harcourt Brace US$2.8 million per year to develop study guides for the same test (Walt 1999). In addition to developing tests that align with a state's content standards, publishers still provide states with off-the-shelf standardized tests. For example, while Massachusetts has its own test aligned with its content standards (i.e. the Massachusetts Comprehensive Assessment System) for grades 4, 8 and 10, it also uses an off-the-shelf test (the Stanford 9) to assess student reading in grade 3. It should also be borne in mind that many local educational authorities still maintain a separate programme of testing for students in their district, e.g. while Massachusetts uses the Stanford 9 test to assess grade 3 reading only, the city of Boston uses the Stanford 9 to assess student achievement at several grades and for several subjects.

The move in the US towards standards-based assessments was accompanied by the rise of the alternative assessment movement. From the late 1980s on, many educators began to move away from the multiple-choice question as the testing format of choice and towards 'alternative' forms of assessment—including open-ended questions, essays, portfolios and per-

Downloaded by [171.66.48.206] at 14:25 23 April 2013

formance tasks. According to a recent report of the Council of Chief State School Officers (CCSSO) (1998), the greatest changes have taken place in the 1990s as more and more states incorporated open-ended and performance exercises into their tests and moved away from reliance on only multiple-choice items. Proponents of the new types of assessments believe that this more open-ended, complex way of assessing students' knowledge can defeat negative test-preparation effects associated with multiple-choice tests, give teachers clear models of acceptable outcomes, measure higher-order skills, and lay bare examinees' thinking processes (Wiggins 1989, 1993). However, as test-developers and users have found out, such items also have a number of problems associated with their use in large-scale district or state high-stakes testing programmes, particularly when they are given to all students at a given grade-level. These problems include inefficiency, administrative inconvenience, subjectivity and bias in scoring, difficulties in standardizing conditions of support for teachers administering the tests within a school and for the actual administration itself, lack of comparability of results, poor generalizability because of the small number of items that can be asked, and high cost. Such problems plagued Kentucky in the 1990s when it introduced its state-mandated, large-scale assessment known as the Kentucky Instructional Results Information System. Kentucky had contracted with an outside test-developer, Advanced Systems, to develop this test which was based on the state's content standards. The assessments in each subject area for the 1992–1993 to 1995–1996 assessment cycles were primarily comprised of open-ended items and portfolio assessments (Kentucky Department of Education 1996). However, because of many of the aforementioned problems, the state had to make substantial changes for the 1997–1998 assessments, including removing test-day performance exercises and adding multiple-choice questions (*Kentucky Post* 1997, Reidy 1997). England and Wales provide another example of a system that initially embraced the use of high-stakes performance assessments as part of its national curriculum. However, difficulties encountered with these innovative assessments eventually led to a return to more traditional modes of testing (Dearing 1993, Thomas *et al.* 1995).

The emergence of computer technology and the possibilities it presents for the design, administration and scoring of tests has had perhaps the greatest impact on the testing industry in recent times. From around the early 1980s, computers began to move from being part of the technical hardware that supported data analysis and test-score reporting to being an integral part of the test development and administration process. Innovations such as computer-based testing (CBT) and computer adaptive testing (CAT) meant that computers became part of the very process of testing itself. CBT differs from conventional testing primarily in the fact that students answer questions using a computer rather than pencil and paper. CAT takes more advantage of computer technology by adapting the test to suit the ability or knowledge level of the test-taker, i.e. the computer selects questions based in part on previous responses, tailoring the test to individual skill levels. CBT and CAT have also brought more flexibility to the overall test-administration process. For example, depending on the testing programme, individuals can register by e-mail or telephone, pay by credit

Downloaded by [171.66.48.206] at 14:25 23 April 2013

card, test by appointment in a designated centre, and receive scores at the end of the session. Testing organizations can electronically exchange questions and examinee responses with test centres, and send scores to institutions in a similar fashion.

The commercial testing industry in general, and Educational Testing Service (ETS) in particular, has noted the potential of CBT. In 1992, ETS computerized the *Graduate Record Examination*, and this was followed by computerized versions of the *SAT1: Reasoning Test* and the *Test of English as a Foreign Language*, among others. By the 1997–1998 academic year, about a million examinees took computerized tests (Bennett 1999). Although this is a relatively small number in comparison to the overall number of tests taken during this period, the numbers will most likely increase sharply over the next few years—especially in light of the fact that the *Graduate Record Examination General Test* is now only offered on computer. However, a possible impediment to such growth is the high costs being faced by testing companies in the area of item development. Testing companies have found that they require much larger item pools in order to address the new security and technical issues that CAT has introduced. Cost issues are also proving to be an impediment to the need to incorporate recent findings from psychometrics and cognitive science into test design (Office of Technology Assessment 1992: 243). In regard to the latter, it is important to note that the air of technological sophistication that surrounds CBT and CAT has not necessarily had a significant effect on what the test actually measures. As Bennett (1999: 3) points out:

> Like many innovations in their early stages, today's computerized tests automate an existing process without reconceptualizing it to realize the dramatic improvements that the innovation could allow. Thus, these tests are substantively the same as those administered on paper; they measure the same skills, use the same behavioral designs, and depend primarily on the same types of tasks.

The next generation of large-scale electronic tests, we hope, will steadily incorporate advances in technology, psychometrics, and, in particular, cognitive science in order to capitalize on the potential of this testing format.

There is little doubt that technical developments such as CBT and CAT will continue to contribute to the rise of standardized testing in the next century. However, these developments also raise serious questions about educational opportunity. For example, is it fair to assess students on computer when they do not have access to a computer in their everyday lives? In a recent report on technology and schooling, the newspaper *Education Week* (1998) notes that among US grade 8 students, 57% reported never or hardly ever using a computer when they do mathematics in school. In addition, 37%of grade 8 students reported that there was no computer at home, and another 15%who did have a computer at home said that they never or hardly ever used that computer for school work. The significant expenses associated with item development for CAT also raise questions about opportunity and access issues. If the high costs involved in test-production are passed along to the consumer (i.e. the student), there

Downloaded by [171.66.48.206] at 14:25 23 April 2013

are implications in terms of whether certain students can afford to take these types of tests and whether they can afford to take them more than once if they need to improve their scores. Alternatively, if the costs are absorbed by the testing industry, how will this effect the trend towards consolidation and takeover by larger corporations that has defined the industry to date? These are important questions and ones that warrant a more detailed discussion than we have space for in this paper. Instead, we now change direction slightly and view the testing industry primarily in terms of quantitative measures of its commercial growth and expansion over time. The next section discusses some indirect indicators that allow us to track the expansion of the commercial enterprise of testing over the last 100 years.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Commercial developments

It has long been recognized that a relatively small number of testing companies now account for the bulk of the test sales in the US Elhi market (Hoffman 1962, Buros 1974a, Fremer 1989). Haney *et al.* (1993) identified seven major testing companies, each with estimated annual gross revenues, mainly from the testing business, of US$15 million or more: ETS, NCS, California Test Bureau, PsychCorp, ACT, Riverside Press, and Scantron. ETS (under contract from the College Board) and ACT control the college entrance examination market, NCS and Scantron are the major players among suppliers of scoring services and machines, and PsychCorp and Riverside Press publish some of the most popular Elhi achievement tests, including the *Stanford Achievement Tests*, the *Metropolitan Achievement Test* (the 'Metropolitan'), the *Iowa Tests of Basic Skills*, and the *Iowa Tests of Educational Development*.

In order to sketch the dimensions of this testing marketplace, we consider a variety of indirect indicators of the extent of the market for tests. It is difficult to obtain exact figures on the scale and volume of this testing marketplace: this is in part due to the fact that because so many different agencies and people administer tests, it is impossible to track down all of them; it is also, in part, attributable to the secretive nature of the testing industry itself. Given the paucity of evidence available on the volume of testing over time, we examined five indirect indicators of growth in testing:

- recent increases in state-mandated testing programmes,
- aggregate sales of tests,
- revenues of four testing companies,
- the rise in price of test booklets and scoring services, and
- references to testing in the education literature.

Recent increases in state-mandated testing programmes

Figure 1 presented a graphical representation of the growth in numbers of states authorizing minimum-competency tests and assessment programmes from 1960–1985. As discussed earlier, the number of state-mandated assessment programmes rose quickly during this period—from one in 1960 to 35 by 1985. The increase in minimum competency testing at the state level was even more dramatic—rising from one such programme in 1965 to 34 by 1976. Figure 2 illustrates the current trend towards standards-based state assessments.

Since the American Federation of Teachers (AFT) began tracking the rise in standards-based assessments in 1995, the number of states using this form of assessment has risen rapidly. In fact, the latest AFT report (1998) shows that the number of states claiming that they use or plan to use standards-based assessments has increased from 33 in 1995 to 47 in 1998—about a 40% increase. These figures are somewhat inflated as they include states that do not yet have standards-based assessments in place; however, the figures are consistent with overall trends noted by other organizations such as the CCSSO. Thus the recent CCSSO report (1998) notes that most US states now administer standardized student assessments once per year in selected subjects and specific grades to all students, and that many states are using both their own tests as well as off-the-shelf tests. For example, in order to assess achievement in English, mathematics, science and social studies, the state of Georgia uses the commercially produced *Iowa Tests of*

Downloaded by [171.66.48.206] at 14:25 23 April 2013

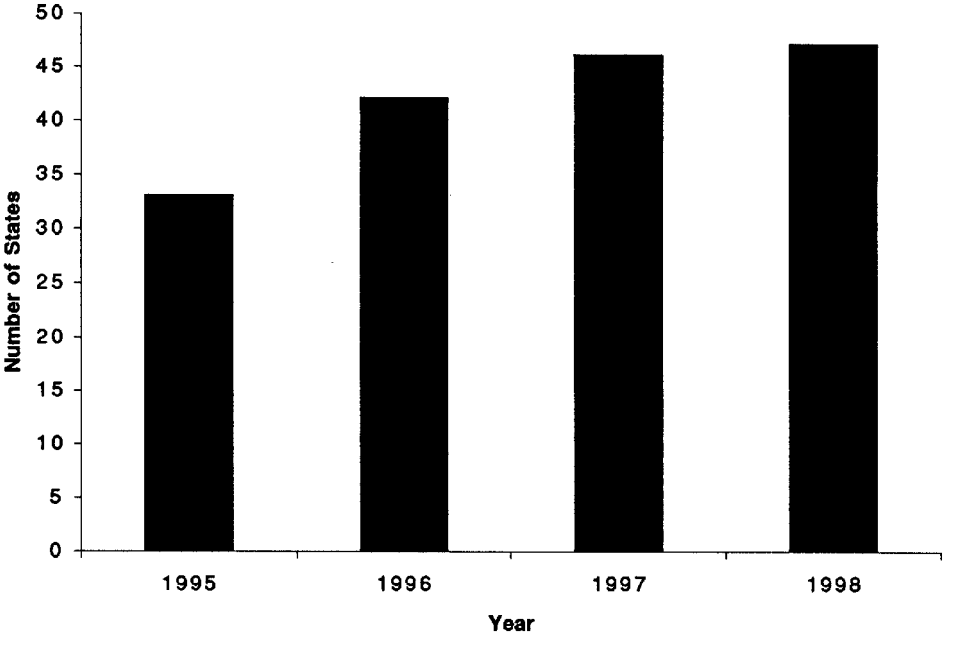

Figure 2. Number of states reporting that they have or plan to have assessments aligned with their standards, 1995–1998 (Source: American Federation of Teachers 1998).

Basic Skills at grades 3, 5 and 8 and its own high school graduation tests in grades 11 and 12.

Aggregate sales of tests

Data on the dollar volume of sales of standardized tests for the US Elhi market has been available for several decades from the *Bowker Annual of Library and Book Trade Information* (R. R. Bowker Company 1970–1998). The *Bowker Annual* gets its sales figures from the Association of American Publishers' (AAP) *Industry Statistics Report*. Figure 3 shows the reported sales figures for standardized tests for the Elhi market for 1955 through 1997, as reported in the *Bowker Annual*. In order to adjust for inflation all costs were converted to constant 1998 dollars (using the yearly consumer price index as the basis for adjustments).

Figure 3 shows a dramatic growth in test sales over the last four decades from less than US$7 million in 1955 to over US$263 million in 1997. In the last eight years alone, test sales increased by about 50%or US$88 million. Although this is a substantial growth in test sales, even after adjusting for inflation, it should be noted that Elhi test sales figure for 1997, as reported by AAP, still represents only a small percentage of the US$22 billion in total book publishing sales in the same year.

Because most publishers treat their sales figures as proprietary, Elhi test sales figures cannot be disaggregated by publisher. Further, annual reports to stockholders by parent companies usually do not break out sales figures for their testing subsidiaries. Hence, we could find no way of checking the

Downloaded by [171.66.48.206] at 14:25 23 April 2013

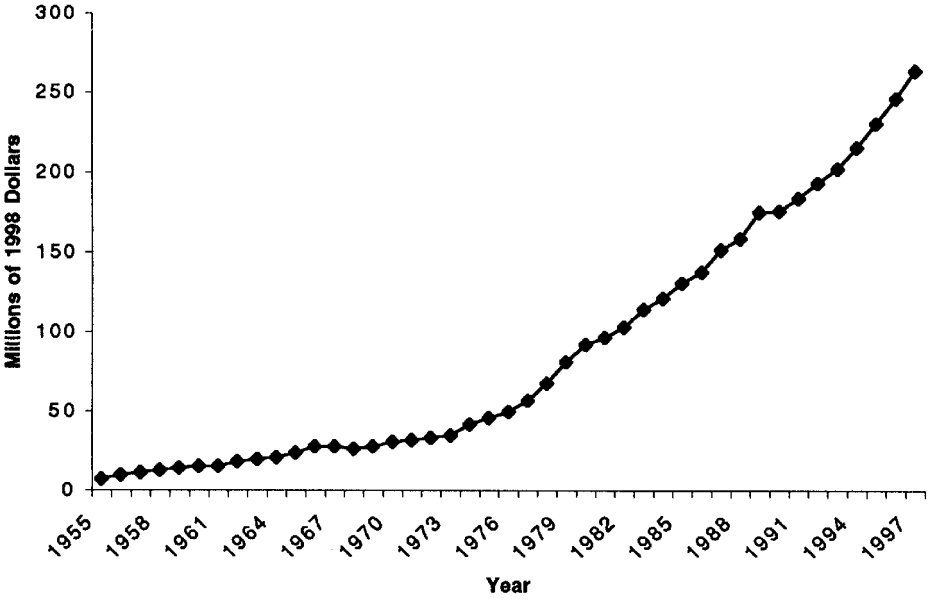

Figure 3. Standardized test sales, 1955–1997, in millions of 1998 US dollars (Source: R. R. Bowker Company 1970–1998).

accuracy of AAP data reported in the *Bowker Annual*. However, we believe that these sales figures may be incomplete in several regards. In particular, AAP data reported in the *Bowker Annual* do not cover three aspects of the testing that affects US elementary and secondary education. First, they do not encompass some test-scoring services (for example the revenues of NCS, which has a dominant role in the test-scoring market). Moreover, they do not include sales figures from companies that are not publishers of tests *per se* but that build standardized Elhi achievement tests on contract for states and districts (e.g. Advanced Systems and National Evaluation Systems). Although there is a relatively small number of these companies, they do a significant business—particularly in states with statewide testing programmes. Also missing from AAP estimates are revenues of ACT and ETS. While these two firms are not typically viewed as part of the Elhi testing market, their college admissions tests represent an important portion of the testing experienced by secondary students in the USA.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Revenues of four testing companies

Although comprehensive data on the dollar volume of test sales are not available, we have been able to locate data on revenues of four of the major testing companies over various periods between 1970 and 1998: namely, for ETS between 1970 and 1998; NCS between 1980 and 1998; ACT between 1972 and 1998; and, Scantron between 1980 and 1998. Figure 4 shows the revenue trends for these four firms over these periods. All amounts are in 1998 dollars.

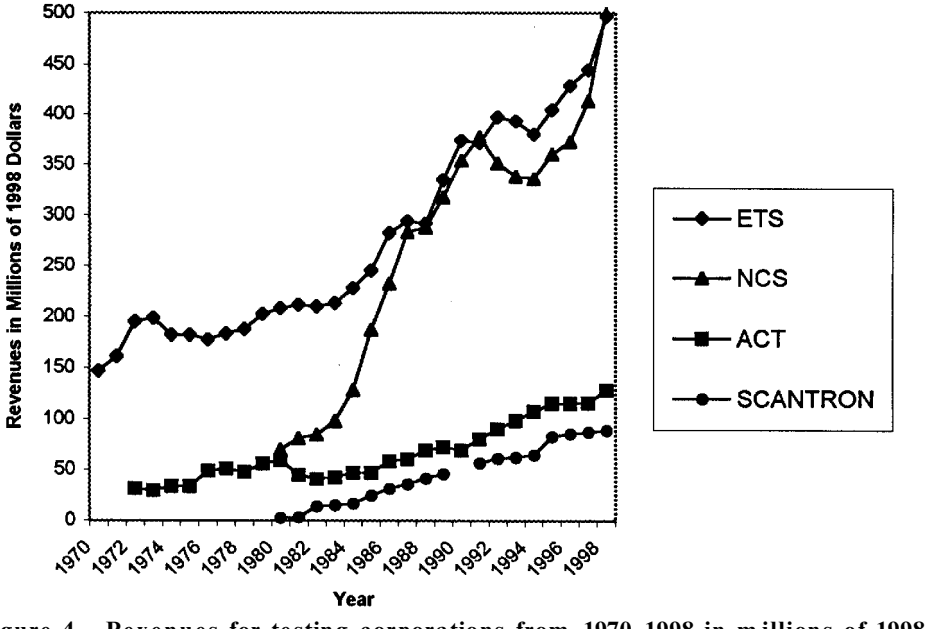

Figure 4. Revenues for testing corporations from 1970–1998 in millions of 1998 US dollars (Source: Company annual reports).

As indicated in figure 4, the total revenues of ETS show a dramatic increase from around US$147 million in 1970 to over US$496 million in 1998, representing a tripling in revenue over the three decades. The total revenues for 1998 also represent an increase of US$52 million over the previous year's revenues with US$45 million of this increase coming from testing activities alone. In fact, over the last 25 years, some 85–90% of ETS total revenues have come from testing services it provides to its clients. An interesting aspect of the ETS figures is that, while ETS is not usually considered a player in Elhi testing, much of its revenues come from the college admissions testing of secondary school students. In 1998, the last year for which we have Elhi sales data from AAP, ETS revenue from College Board testing alone (almost US$157 million, not included in AAP data) was nearly 60% of the value that AAP reported for all Elhi test sales (US$263 million) and represented around 32% of ETS's total revenues for that year.

Revenues of NCS (see figure 4) have increased even more sharply over the last decade, skyrocketing from US$69 million in 1980 to US$505 million in 1998—a rate of increase far outpacing that of ETS. Although we do not have detailed information on the breakdown of NCS total revenues, it appears that the vast majority of NCS revenues—on the order of 80–90%—come from scanning services, test-building and test sales. NCS's emergence as a major player in the testing marketplace illustrates two major trends. Increasingly, the market appears to be fracturing so that, for a given test, different organizations can be involved in the sponsorship, development, administration, scoring, interpretation and use of results. Also, NCS's recent sharp growth—revenues grew by about US$221 million or 78% between 1990 and 1998—indicates the increasing importance of computer technology in the testing marketplace, including the use of computers not just to score test results but also to produce reports 'interpreting' test results.

The ACT Programme was founded in the 1950s as a not-for-profit organization to produce tests for college admissions. ACT still serves colleges but also provides services to K-12 education and educational agencies, business and industry. It also provides a broad range of supplementary materials and services. ACT (see figure 4) has shown more modest revenue growth than some of the other companies over the last 27 years, increasing from US$31 million in 1972 to US$128.3m in 1998. This represents a four-fold increase in revenues over nearly three decades—roughly equivalent to ETS's pace of growth during the same period. However, unlike ETS, ACT actually experienced a period in 1981–1983 when total revenues declined but this downturn did not represent a slump in ACT testing business, rather it reflected ACT's loss of a major federal contract to process college and university student financial aid forms.

Revenues for Scantron, the fourth company for which data is presented in figure 4, have been much smaller than those of ETS, NCS and ACT. This is hardly surprising because Scantron was only founded in 1972, whereas the other companies have existed since 1947, 1962 and 1958 respectively. However, between 1980 and 1988 (when it became a subsidiary of John Harland Co.), Scantron's revenues increased from approxi-

Downloaded by [171.66.48.206] at 14:25 23 April 2013

mately US$2 million to US$41 million, representing an annual rate of growth that considerably outpaced the growth rates of the larger firms. Since 1988, Scantron's revenues have increased from US$41 million to US$88 million. However, in that same time period their core business moved from 90% educational to just 50% educational (with the balance being taken up by an increase in services they provide to the commercial sector). Thus, not all of the growth in profit can be directly attributable to the education market. Scantron illustrates several significant recent trends in the testing marketplace: the increasing importance of computer technology in testing, the importance of the test-scoring market as compared with the more traditional test-publishing market, and the rapid pace of corporate takeovers and reorganizations.

Although we do not have sufficiently detailed information on sources of revenues for each of these four companies to pinpoint trends very precisely, comparing figure 4 with figure 3 does suggest one generally common trend. For the four companies for which we have annual revenue data (ETS, NCS, ACT and Scantron) and also for the test publishers reporting to AAP, the period of the 1980s and 1990s appears to have been one of unprecedented growth in sales for the testing industry.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

The rise in the price of test booklets and scoring services

In order to place data about test sales in perspective, we examined the price of the test booklets and the scoring service for the achievement batteries of the three largest publishers in the US Elhi market—Harcourt Brace Jovanovich/PsychCorp (which publishes the *Metropolitan* and the *Stanford Achievement Test*), California Test Bureau/McGraw-Hill (which publishes the *California Achievement Test*), and Houghton-Mifflin/Riverside Press (which publishes *Iowa Tests of Basic Skills*). We wish to consider the extent to which reported increases in test sales may be due to increases in volume of testing versus increases in prices of tests and testing services sold. Specifically, we examined costs of test booklets, machine-scorable answer sheets and scoring services as reported in the *Mental Measurements Yearbooks* (Buros 1938, 1940, 1949, 1953, 1959, 1965, 1974b, 1978, Conoley and Kramer 1989, 1992, 1995, Mitchell 1985) or in recent catalogues of the three test publishers. In order to adjust for inflation, all costs were converted to constant 1998 dollars. Summary results of these analyses are shown in figures 5 and 6. The answer sheet and scoring costs are averages across the three publishers. The booklet cost figures in adjusted dollars are shown in figure 5;[7] the scoring cost figures in constant 1998 US dollars in figure 6.

Figure 5 shows that the real price of purchasing a test booklet (that is, the constant dollar price) has approximately tripled over the last 60 years, with the Iowa test booklet showing more fluctuations in price than any of the Metropolitan, Stanford or California test booklets. Interestingly, the price of scoring services decreased from 1972 to 1985, but has increased since then. However, when the increase in costs for both test booklets and scoring services (figure 5 and figure 6) are compared to the eight-fold

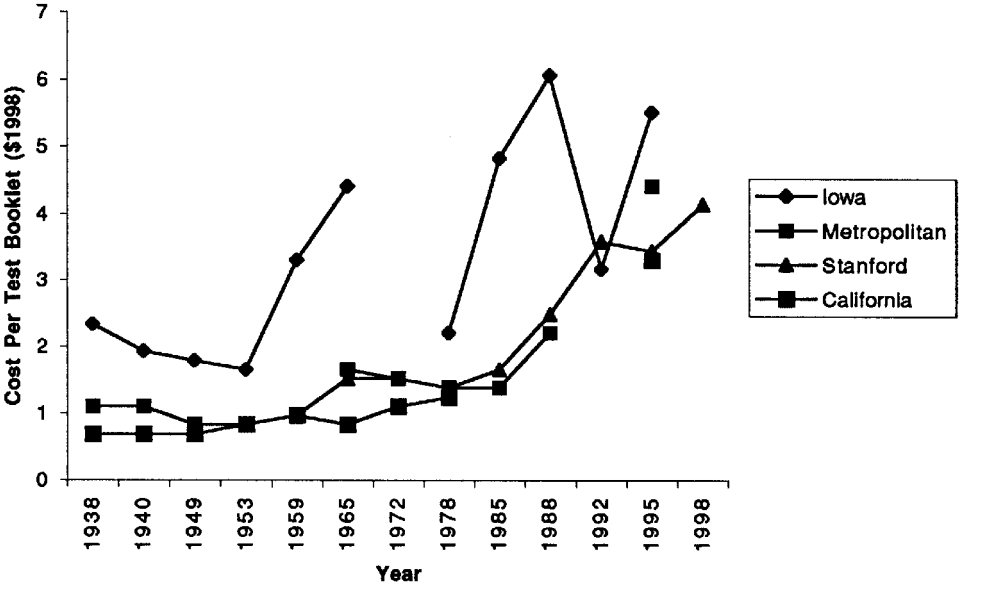

Figure 5. Price per test booklet over time in 1998 US dollars (Sources: Mental Measurements Yearbooks and company catalogues).

Downloaded by [171.66.48.206] at 14:25 23 April 2013

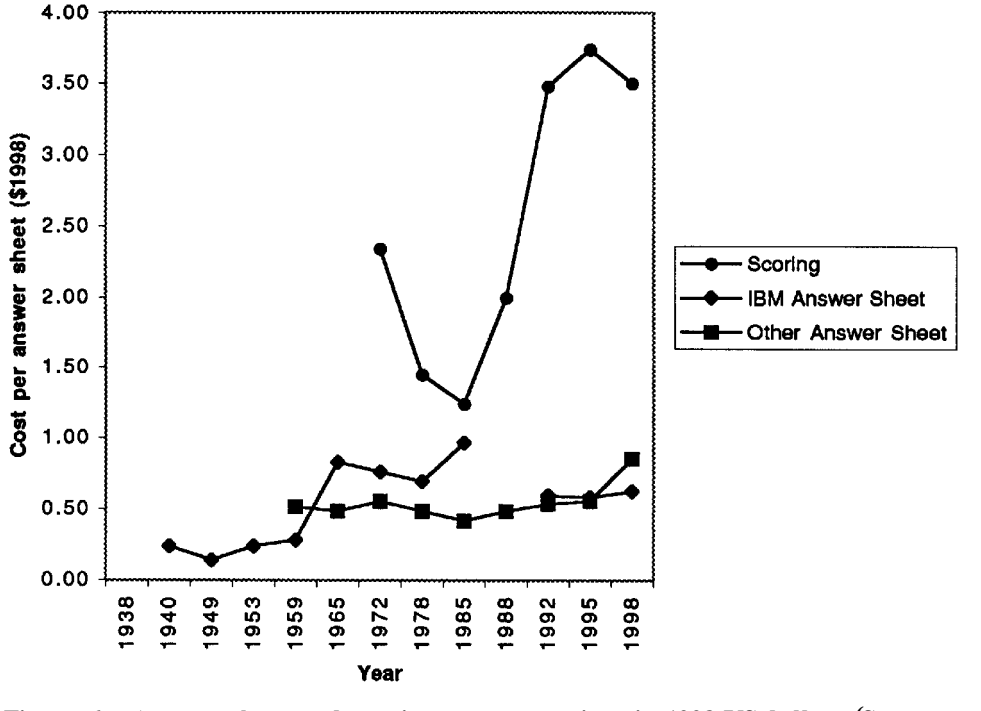

Figure 6. Answer sheet and scoring costs over time in 1998 US dollars (Sources: Mental Measurement Yearbooks and company catalogues).

increase in total test revenue during the same period (using figures reported by AAP but adjusting for inflation), it seems clear that the substantial increase in Elhi test sales over the last 25 years cannot be explained solely by increases in the costs of tests and related services. Instead, the increase is due in large measure to increases in the volume of testing.

References to testing in the education literature

An indirect indicator, developed by Haney (1986), documents the increased attention over the decades to testing's importance in the educational realm (Haney *et al.* 1993). To show growth in the volume of testing over time, he charted the number of citations under the rubric 'testing' (as indicated by the number of column-inches) from 1932–1985 in the *Education Index* (H. W. Wilson 1932–1998). For comparative purposes, and because he argued curriculum issues should be a central focus of schooling, the number of citations under 'curriculum' were also charted.[8] The data shown in figure 7 are updated through 1998.

Figure 7 shows that the average annual number of column-inches devoted to citations concerning curriculum has increased only modestly over the last 66 years—from 50 to 100 inches per year in the 1930s and 1940s to around 150 in recent years. In contrast, column-inches devoted to tests and scales have increased greatly, from only 10 to 30 in the 1930s and 1940s to up to 400 in the 1980s. The past few years have seen a decline in the number of citations regarding testing (it currently stands at around

Downloaded by [171.66.48.206] at 14:25 23 April 2013

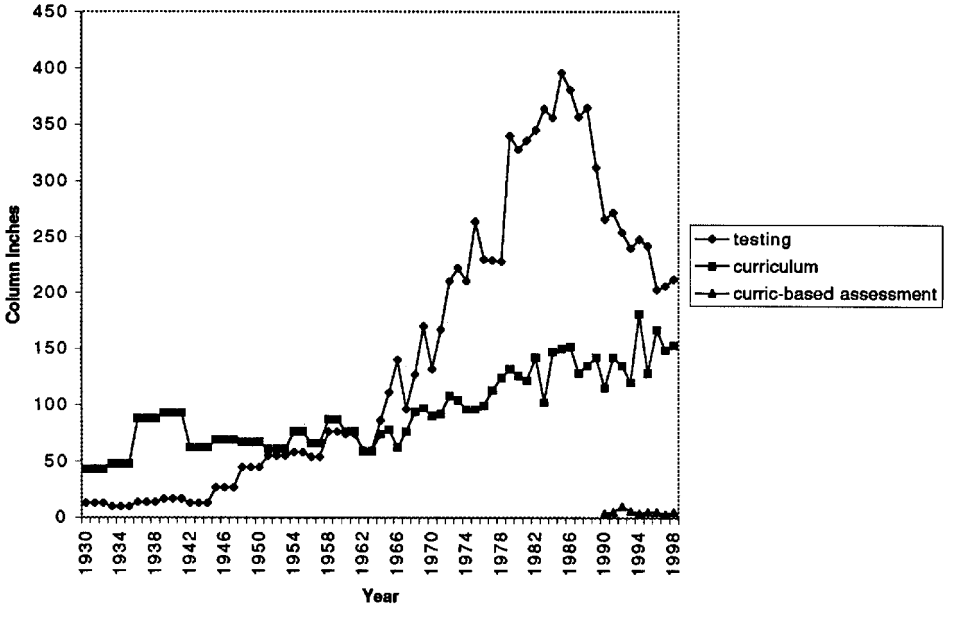

Figure 7. ***Education Index*** **listings under 'testing', 'curriculum' and 'curriculum-based assessment', 1932–1998 (Sources: Haney *et al.* 1993, H. W. Wilson 1932–1998, Madaus *et al.* 1997).**

210); however, the new rubric 'performance-based assessment' was added to the *Education Index* in 1992 to reflect prevailing testing terminology, and those citations are not included in our update. We do include data in figure 7 for 'curriculum-based assessment', another category that was implemented in 1990. The inclusion of this term is testimony to the trend towards standards-based assessments at the state level, which we have commented on earlier. Although these indices are admittedly crude, the data certainly highlight the prominence of testing in the educational literature, particularly since the mid-1960s.

Conclusion

The history of testing in the USA and elsewhere throughout the 20th century shows that most changes or developments in the technology of testing have been directed at making testing more efficient, manageable, standardized, objective, easier to administer, and less costly in the face of increasing numbers of examinees. Efficiency, management and cost concerns are issues for any industry—not just the testing industry. However, these concerns should not be the sole criteria by which the enterprise of testing, and its success, is evaluated; there is a need to evaluate testing in terms of whether a particular test use is appropriate, and use for which the test was designed. The need to evaluate test use in terms of its 'appropriateness' is particularly important in light of the fact that tests are increasingly being used by policymakers and others to make high-stakes decisions about students, teachers, schools and districts (NRC 1999).

In many other areas where technology and policy intersect, the public insists on oversight—including technical oversight—to protect individuals from unintended negative effects. For example, faced with the policy decision to introduce a major new untried medical technology to millions of children, particularly a treatment that would be given to healthy children as well those who were ill, the US public would ask about the safety, efficacy, quality, and social and economic effects of the new technology or treatment; and public agencies have been established to address such concerns systematically. The effects of testing are now so diverse, widespread and serious in the US that we believe it is necessary to establish similar mechanisms for catalysing inquiry about and systematic independent scrutiny of them.

However, for most of this century, there has been no infrastructure for independently evaluating a testing programme before or after implementation, or for monitoring test use and impact. The commercial testing industry does not as yet have any structure in place for the regulation and monitoring of appropriate test use. In addition, although there has been some development of standards and codes of testing by professional organizations (e.g. American Educational Research Association *et al.* 1985, Joint Committee on Testing Practices 1988), there is no parallel mechanism to ensure that they are implemented.

What is needed is an ongoing monitoring of the testing enterprise—a co-ordinated effort to ensure that tests that are developed are technically

Downloaded by [171.66.48.206] at 14:25 23 April 2013

sound and appropriately used, and that the consequences of this use are monitored in order to detect and remove any unintended negative consequences. Effective 1 September 1998, such an organization came into existence. The National Board on Educational Testing and Public Policy—a monitoring body based at Boston College—was reconstituted to provide a venue for evaluating testing programmes with an emphasis on formative evaluation. With initial funding from the Ford Foundation, the Board's overall goal is to encourage test-makers, policy makers, and consumers to use tests appropriately and responsibly and to improve the balance of benefits to harms associated with testing programmes. Such a monitoring body is not at odds with the efficiency and costs concerns that have characterized the testing enterprise for much of this century. Rather, it is a mechanism that will allow both the producer (i.e. the test publisher) and the consumer (i.e. the person who buys, or uses the test) to discover problems before they become costly mistakes and to implement testing programmes that have the best chance of being both commercially successful and educationally beneficial.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

Notes

1. Sections of this paper are based on previous work by the National Board on Educational Testing and Public Policy.
2. Included in this section is a discussion of some developments in test theory which, while not being strictly technical innovations, either underpin the technical innovation or could only be implemented because of the advent of a particular technical innovation.
3. For a detailed discussion of testing as a technology, see Madaus (1994). For a general discussion of testing through the ages, see Madaus and O'Dwyer (1999).
4. A survey of 200 school superintendents by Haggerty in 1918 revealed that tests were being used to bring about six different types of change in schooling. These were changes in classification of pupils, school organization, courses of study, time devoted to subject, methods of instruction, and methods of supervision.
5. Bloom *et al.*'s (1956) *Taxonomy* contains six categories purported to be hierarchical: knowledge, comprehension, application, analysis, synthesis, and evaluation. The *Taxonomy* is currently under revision.
6. With the merger of Harcourt Brace Jovanovich and World Book in 1962 and Harcourt Brace Jovanovich's acquisition of PsychCorp in 1969, the former World Book tests became part of PsychCorp's line of products.
7. Although the Iowa costs appear much larger, the Iowa booklet bundles the test batteries for all the grades. The other two publishers market the booklets separately by grade level.
8. Figure 7 was constructed by measuring the number of column-inches devoted to lines concerning testing and curriculum in every volume of the *Education Index* from 1932 through 1998. Over these volumes there were some changes in the index rubrics concerning testing and curriculum. (See Haney *et al.* (1993) and Madaus and Raczek (1995).)

References

American Educational Research Association, American Psychological Association, National and Council on Measurement in Education (1985) *Standards for Educational and Psychological Testing* (Washington, DC: American Psychological Association).

AMERICAN FEDERATION OF TEACHERS (1998) *Making Standards Matter 1998: An Annual Fifty-State Report on Efforts to Raise Academic Standards* (Washington, DC: American Federation of Teachers).

BAKER, F. (1989) Computer technology in test construction and processing. In R. Linn (ed.), *Educational Measurement*, 3rd edn (New York: Macmillan), 409–428.

BENNETT, R. E. (1999) *Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing* (Princeton, NJ: Educational Testing Service).

BLOOM, B. S., ENGELHART, M. D., FURST, E. J., HILL, W. H. and KRATHWOHL, D. R. (1956) *Taxonomy of Educational Objectives Handbook I: The Cognitive Domain (*New York: David McKay).

BUROS, O. K. (ed.) (1938) *The Nineteen Thirty Eight Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1940) *The Nineteen Forty Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1949) *The Third Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1953) *The Fourth Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1959) *The Fifth Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1965) *The Sixth Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1974a) *Tests in Print II: An Index to Tests, Test Reviews, and the Literature on Specific Tests* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1974b) *The Seventh Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

BUROS, O. K. (ed.) (1978) *The Eighth Mental Measurements Yearbook* (Highland Park, NJ: Gryphon Press).

CHAPMAN, P. D. (1980) Schools as sorters: Lewis M. Terman and the intelligence testing movement, 1890–1930. *Dissertation Abstracts International*, 40. University Microfilms No. 80-11615.

COLEMAN, J. S., CAMPBELL, E. Q., HOBSON, C. J., MCPARTLAND, J., MOOD, A. M. and WEINFIELD, F. D. (1966) *Equality of Educational Opportunity* (Washington, DC: U.S. Government Printing Office).

CONOLEY, J. C. and KRAMER, J. L. (eds) (1989) *The Tenth Mental Measurements Yearbook* (Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln).

CONOLEY, J. C. and KRAMER, J. L. (eds) (1992) *The Eleventh Mental Measurements Yearbook* (Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln).

CONOLEY, J. C. and KRAMER, J. L. (eds) (1995) *The Twelfth Mental Measurements Yearbook* (Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln).

COUNCIL OF CHIEF STATE SCHOOL OFFICERS (1998) *Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance: A 50-State Report, December 1998* (Washington, DC: Council of Chief State School Officers).

DEARING, R. (1993) *The National Curriculum and its Assessment: An Interim Report* (Hayes, UK: School Examinations and Assessment Council, National Curriculum Council).

EDUCATION INDEX, (1932–1998) (New York: H. W. Wilson).

EDUCATION WEEK (1998) *Technology Counts' 98* (Bethesda, MD: Education Week/Milliken Exchange on Education Technology).

FREMER, J. J. (1989) Testing companies, trends, and policy issues: a current view from the testing industry. In B. Gifford (ed.), *Testing Policy and the Politics of Opportunity Allocation: The Workplace and the Law* (Boston: Kluwer), 61–80.

HAGGERTY, M. E. (1918) Specific uses of measurement in the solution of school problems. In G. M. Whipple (ed.), *The Measurement of Educational Products*, 17th Yearbook, Part 2, of the National Society for the Study of Education (Bloomington, IL: Public School Publishing Co.), 25–40.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

HANEY, W. (1984) Testing reasoning and reasoning about testing. *Review of Educational Research*, 54(4), 597–654.

HANEY, W. (1986). College admissions testing and high school curriculum: uncertain connections and future directions. In *Measures in the College Admissions Process: A College Board Colloquium* (New York: College Entrance Examination Board), 32–52.

HANEY, W. M., MADAUS, G. F. and LYONS, R. (1993) *The Fractured Marketplace For Standardized Testing* (Norwell, MA: Kluwer).

HOFFMAN, B. (1962) *The Tyranny of Testing* (New York: Crowell-Collier).

HOLMEN, M. G. and DOCTER, R. (1972) *Educational and Psychological Testing: A Study of the Industry and its Practices* (New York: Russell Sage Foundation).

HORACE MANN PAPERS, Box 8 (1845–1846) (Boston, MA: Massachusetts Historical Society).

JOINT COMMITTEE ON TESTING PRACTICES (1988) *Code of Fair Testing Practices in Education* (Washington, DC: National Council on Measurement in Education).

JUDD, C. H. (1918) A look forward. In G. M. Whipple (ed.), *The Measurement of Educational Products*, 17th Yearbook, Part 2, of the National Society for the Study of Education, Part II (Bloomington, IL: Public School Publishing Co.), 152–160.

KANIGEL, R. (1997) *The One Best Way: Frederick Winslow Taylor and the Enigma of Efficiency* (New York: Viking).

KENTUCKY DEPARTMENT OF EDUCATION (1996) *Kentucky School and District Accountability Results* (Frankfurt, KY: Kentucky Department of Education).

KENTUCKY POST (1997) Editorial. *Kentucky Post*, 16 April, 10.

MADAUS, G. F. (1993) A national testing system: manna from above: an historical/technological perspective. *Educational Assessment*, 1(1), 9–26.

MADAUS, G. F. (1994) A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review*, 64(1), 76–95.

MADAUS, G. F. (1995) A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. In M. T. Nettles and A. L. Nettles (eds), *Equity and Excellence in Educational Testing and Assessment* (Boston: Kluwer), 23–68.

MADAUS, G. F. and O'DWYER, L. M. (1999) A short history of performance assessment. *Phi Delta Kappan*, 80(9), 688–695.

MADAUS, G. F. and RACZEK, A. E. (1995) The extent and growth of educational testing in the United States: 1956–1994. In H. Goldstein and T. Lewis (eds), *Assessment: Problems, Developments and Statistical Issues* (Chichester, UK: Wiley), 145–165.

MADAUS, G. F., CLARKE, M. and O'LEARY, M. (in press) A century of standardized mathematics testing. In G. M. Stanic and J. Kilpatrick (eds), *A Recent History of Mathematics Education in the United States and Canada* (Reston, VA: National Council of Teachers of Mathematics).

MADAUS, G. F., RACZEK, A. and CLARKE, M. M. (1997) The historical and policy foundations of the assessment movement. In A. L. Goodwin (ed.), *Assessment for Equity and Inclusion: Embracing all our Children* (New York: Routledge), 1–33.

MAGER, R. F. (1962) *Preparing Instructional Objectives* (Belmont, CA: Fearon).

MITCHELL, J. V. (ed.) (1985) *The Ninth Mental Measurements Yearbook* (Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln).

NATIONAL COMMISSION ON EXCELLENCE IN EDUCATION (1983) *A Nation at Risk: The Imperatives for Educational Reform* (Washington, DC: US Department of Education).

NATIONAL COMMISSION ON TESTING AND PUBLIC POLICY (1990) *From Gatekeeper to Gateway: Transforming Testing in America* (Chestnut Hill, MA: National Commission on Testing and Public Policy).

NATIONAL RESEARCH COUNCIL (1999) *High Stakes: Testing for Tracking, Promotion and Graduation* (Washington, DC: National Academy Press).

OFFICE OF TECHNOLOGY ASSESSMENT (1987) *State Educational Testing Practices: Background Paper* (Washington, DC: Science, Education and Transportation Program, Office of Technology Assessment).

OFFICE OF TECHNOLOGY ASSESSMENT (1992) *Testing in American Schools: Asking the Right Questions* (Washington, DC: US Government Printing Office), OTA-SET-519.

Downloaded by [171.66.48.206] at 14:25 23 April 2013

POPHAM, W. J. (1978) *Criterion-referenced Measurement* (Englewood Cliffs, NJ: Prentice-Hall).

PORTER, T. M. (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, NJ: Princeton University Press).

POSTMAN, N. (1992) *Technopoly: The Surrender of Culture to Technology* (New York: Knopf).

PRESSEY, S. L. and PRESSEY, L. C. (1922) *Introduction to the Use of Standardized Tests* (Yonkers, NY: World Book Company).

R. R. BOWKER COMPANY (1970–1998) *The Bowker Annual of Library and Book Trade Information*, Vols. 16–42 (New York: R. R. Bowker).

REIDY, E. (1997) Personal communication.

SAMELSON, F. (1987) 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? In M. M. Sokal (ed.), *Psychological Testing and American Society, 1890–1930* (New Brunswick, NJ: Rutgers University Press), 113–127.

THOMAS, S., RACZEK, A. E. and MADAUS, G. F. (1995) *Differential Impact of Performance Assessment: The British Experience* (Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Public Policy, Boston College).

TYLER, R. W. (1934) *Constructing Achievement Tests* (Columbus, OH: Ohio State University, Bureau of Educational Research).

UNITED KINGDOM. PARLIAMENT (1886) *Report of the Royal Commission to Inquire into the Workings of the Elementary Education Acts (England and Wales)*, Vol. 1, Minutes of Evidence. C. 4863.

US DEPARTMENT OF EDUCATION (1991) *America 2000: An Education Strategy: Sourcebook* (Washington, DC: U.S. Department of Education).

WALT, K. (1999) TAAS troubles: publisher taking precautions as it faces possible costly loss. *Houston Chronicle*, 2 May, 15.

WIGGINS, G. (1989) A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*, 70(9), 703–713.

WIGGINS, G. (1993) Assessment: authenticity, context, and validity. *Phi Delta Kappan,* 75(3), 200–214.

WISE, A. E. (1979) *Legislated Learning: The Bureaucratization of the American Classroom* (Berkeley, CA: University of California Press).

Downloaded by [171.66.48.206] at 14:25 23 April 2013