Part One

Setting the Stage

Chapter 1

Examinations for Educational Productivity

Herbert J. Walberg

This chapter addresses three questions: (1) Where do U.S. schools stand on international examinations relative to those in other affluent countries? (2) Why do they do so poorly at such great cost? (3) How can examinations help? I argue that objective examinations, though imperfect, are reasonable measures of important results of schooling. Though they may not tell the whole story, they can be readily employed to discover effective practices, to improve accountability, and to evaluate choice experiments. Other examinations, such as portfolios and laboratory exercises, are appropriate for assessing students' classroom work but have proven costly and impractical for evaluating schools and districts. In any case, it is important to employ value-added measures to assess the contributions of schools, programs, and staff.

Examinations can be keys to improving the productivity of U.S. schools. Educational policy makers can employ them to evaluate educational organizations, policies, and programs to determine which are most effective and efficient. Value-added

analyses of examination results offer a way to achieve their policy and accountability purposes.

In a free society, however, consumer choice would seem to offer the ultimate and best accountability. Since private and public scholarships are unlikely to predominate soon, examinations can serve to help evaluate various means and degrees of enlarging choice and competition in the educational systems.

Where Do U.S. Schools Stand?

It is increasingly well-known that our secondary school students score poorly on objective examinations compared with those in other economically advanced countries. By such standards, our high school students have long done poorly in these subjects, although primary school students have scored nearer to the average. These differences suggest that our students make poor progress during the school years. But how much worse is their progress relative to that of students in other countries? My report for the Thomas B. Fordham Foundation took up this question and compiled all recent achievement comparisons.[1]

The report compared advanced countries that are members of the Organization for Economic Cooperation and Development in North America, the Pacific Rim, and Western Europe. Among schools in OECD countries, those in the United States made the smallest achievement gains. The longer U.S. students were in school, the further they fell behind students in the other countries. Yet per-student expenditures on U.S. schools are among the very highest. More specifically:

1. In reading, science, and mathematics through eighth grade, U.S. schools ranked last in four of five comparisons of achievement progress. In the fifth case, they ranked second to last.
2. Between eighth grade and the final year of secondary school, U.S. schools slipped further behind those in other countries.

3. Because they made the least progress, U.S. secondary schools ranked last in mathematics attainment and second to last in science—far from the goal to be first in the world by the year 2000, set by the fifty governors and endorsed by Congress and the 1996 presidential candidates.
4. U.S. per-student spending (adjusted for purchasing power) on primary and secondary schools was third-highest among more than twenty advanced countries.
5. Unlike in the past, more secondary school students remain in school on average in comparable countries than in the United States. Thus, their superior gains do not depend merely on student selectivity or higher dropout rates.
6. Because they made the poorest progress and ranked in the highest category of expenditures, U.S. schools, by internationally agreed-upon standards, are the least productive among those in comparable economically advanced countries.

Value-Added Comparisons

These conclusions are based on the most recent, largest, and most rigorous international achievement surveys. Unlike other reports, the conclusions concern the value added largely by schools as indexed by progress made by students during the school years.

Value-added scores are particularly important in evaluating schools. Consider the case of reading. Until children start school at about age six, families, media, and other agencies—rather than schools—are the chief sources of influence on vocabulary and comprehension. For this reason, children start school with widely varying degrees of preparation. Some parents but not others, for example, teach their children to read before first grade. The big education question is: How much progress do students make after they start school?

Static comparisons of schools (employed in the past) are less useful for this purpose because students' tests scores are partly

determined by their experiences before they begin school, attributable to parental efforts, socioeconomic status, and related factors. Thus, gains in achievement during the school years are better indexes of schools' contributions to learning than scores at a single point in time.

Gains, progress, and value added—terms used synonymously here—are particularly important for policy. They allow predictions of eventual attainments. Policies that do not add satisfactory value may be revised. Units of the system, such as primary and secondary schools, may be separately evaluated by measuring students' progress while under their responsibility. In addition, many economists, psychologists, and others believe incentives influence performance. For this reason, principals should give merit raises for recent progress rather than for degrees and years of experience, for which most teachers are paid. If carrots and sticks were employed in education, value-added progress would be one reasonable indicator of teaching merit.

Educational policy makers increasingly recognize the usefulness of value-added indicators. Internationally, the OECD pioneered the use of value-added indicators in the 1995 edition of *Education at a Glance*[2] and has employed them in subsequent reports. Similarly, Dallas, Texas, and Tennessee are employing value-added indicators and incentives to increase school productivity. Other cities and states, such as Chicago and Virginia, employ static indicators to assign schools to probation and, in cases of failure to progress, eventual extinction. Such systems identify schools that serve poor children but that are not ineffective. A fairer and more efficient evaluation system would employ value-added indicators as at least one consideration in evaluating schools.

Why Do U.S. Schools Do So Poorly?

Several problems appear to account for poor productivity of U.S. public schools. After reviewing these, we can consider how effective practices, better accountability, and enlarged choice together with objective examinations may help solve them.

Lack of State Standards

Unlike many other countries, the U.S. education system has no education ministry nor well-defined national goals, curriculum, or testing system. The U.S. system leaves states largely responsible for providing schools, but states leave varying amounts of discretion to local boards. What is taught in classrooms, in turn, is highly variable even within the same school and district. For these reasons, a teacher in any grade cannot depend on what the teacher in the previous grade has taught. The lack of coordination across grades and subjects is especially harmful to children who move, particularly if they also are poor.[3]

Lack of standards means that state and local boards can hardly assess progress made by districts, schools, and teachers. To the extent that curriculum and goals vary, it is difficult to compare schools, which makes accountability for results nearly impossible.

Centralized Finance and Control

Despite the lack of uniform standards and accountability, the governance and funding of public schools have become more centralized in the last half-century, leading to other kinds of inefficiency. States have increasingly assumed responsibility for educational finance, goals, and operations. They paid ever-larger shares of school costs, but the higher the state's share, the worse the state's achievement, despite vast increases in inflation-adjusted per-student spending. Higher state shares make local school boards and administrators less accountable to local citizens since they need not justify expenditures as carefully.[4] California's tie for last place in recent national reading assessments may be attributable to whole-language teaching and highly centralized state funding rather than the greater local control and accountability afforded by local funding.

Larger state shares also entail increased regulation, reporting, bureaucracy, and distraction from learning. Much energy goes into the question of who governs—the federal government, the state, the local district, the school, or the teacher. It is nearly impossible to affix responsibility for results.

Schools and school districts, moreover, have increasingly consolidated into larger units that achieve less. Over the course of a recent fifty-year period, average school enrollments in the United States multiplied by a factor of five, even though large schools tend to be more bureaucratic, impersonal, and less humane. Large middle and junior high schools tend to departmentalize and employ specialized teachers and ancillary staff who confine themselves to their specialties rather than imparting a broad view of knowledge. The teachers in large, departmentalized schools tend to know their students much less well than teachers who have the same students for most subjects for nearly the whole day.

About a half-century ago, there were 115,000 U.S. school districts; now there are about 15,000, the largest of which tend to be least effective. The reasons for their inefficiency are best seen in New York and other large cities that have up to 900 schools. In such huge districts, school board members can hardly name the schools let alone hold them accountable.[5]

On the other hand, small adjacent public school districts and private schools within districts give rise to incentives that cause all schools to compete and raise their productivity, that is, raise achievement and student retention while lowering costs. Choice plans that allow students to cross school and district boundaries may also prove to increase competition and productivity. Choice among schools, nonetheless, is severely constrained, which helps account for poor U.S. productivity.

Lack of Board Accountability

School boards frequently split into factions. And few members have extensive board, business, or education experience. Often serving limited terms, they seem more interested in personnel and ideological issues than in whether the schools are achieving results. Assessing learning progress, moreover, requires some mastery of educational productivity research, psychometrics, and statistics, just as assessing businesses' progress requires accounting and other skills. Few board members or educational administrators have mastered such skills. Instead, they take up such fads as

Ebonics, whole language, authentic tests, and bilingual education—the success of which remain undemonstrated in randomized experiments or statistically controlled research.

Unaccountable Management

Public schools are government-subsidized quasi-monopolies. They are unchallenged by entrepreneurial leadership and the incentives, efficiency, and consumer appeal provided by market competition. With legislators and school boards often under their thumbs, teachers' unions and administrators can exploit forced-choice customers in service of their interests in minimizing workload and maximizing pay and perquisites.

In particular, teachers' unions—few call them professional associations—have actually done well for their members. In college, education majors typically have scored worst or near worst on ability tests among undergraduate majors. Yet as teachers, they have a 180-day school year—the shortest among teachers in industrialized countries (and much less than the 220 or so days most salaried U.S. professionals normally work). In large cities and elsewhere, according to contract, many teachers are in school only about six hours daily. Some grade papers in the evening, but many professionals take work home. In addition, teachers have little accountability, nearly inviolable tenure, and early and generous pensions that increasingly threaten city and state budgets.

Teachers' unions have done better for themselves than for their members. During the last half-century when membership in private-sector unions declined, teachers' unions increased their membership. They contracted for expensive smaller classes, which do little for learning. With fixed budgets, smaller classes actually mean lower teacher salaries because costs must be spread among more teachers. Thus, smaller classes, which increase the number of teachers, indirectly result in an increase in union membership, central coffers, and legislative influence.

Teachers' unions are understandably acting in their own interests of maximizing their benefits while reducing their efforts. It is school boards and state legislators that have been remiss in

failing to provide effective management, informed stewardship, and accountability to citizens who pay the bills. School boards have hardly been a match for nationally organized unions that can bring to negotiations strong, narrow self-interests, statistical research, and specialized expertise.

Harvard and University of Chicago economists Caroline Hoxby and Samuel Peltzman showed that teachers' union success was associated with worse results for students. Their analyses showed that the sharp rise in teachers' union membership and militancy for the period 1971–1991 not only increased per-student costs dramatically but also increased dropout rates and adversely affected examination scores in the forty-eight states surveyed. As teachers' unions grew in membership, income, and power, they gained greater influence over state legislatures, which, in turn, increasingly usurped local control and left the schools increasingly ineffective and unaccountable to local taxpayers.[6]

Lack of Incentives

American schools provide little incentive for educators and students to attain higher standards. A 1996 Public Agenda national survey of high school students showed that three-fourths believe stiffer examinations and graduation requirements would make students pay more attention to their studies. Three-fourths also said students who have not mastered English should not graduate, and a similar percentage said schools should promote only students who master the material. Almost two-thirds reported they could do much better in school if they tried. Nearly 80 percent said students would learn more if schools made sure they were on time and did their homework. More than 70 percent said schools should require after-school classes for those earning Ds and Fs.[7]

In these respects, many teacher educators differ sharply from students and the public. A 1997 Public Agenda survey of education professors showed that 64 percent think schools should avoid competition. More favored giving grades for team efforts than for individual accomplishments.

Teacher educators also differ from employers and other professions on preferred ways of measuring standards or even employing such measures at all. Many employers use standardized multiple-choice examinations with job candidates. So do selective colleges and graduate and professional schools with candidates for admission. Such examinations are required in law, medicine, and other fields for licensing because they are objective and reliable. In the case of teachers, academic mastery (as indicated by objective examination results and completion of rigorous courses) influences their students' achievement. Yet, 78 percent of teacher educators wanted less reliance on objective examinations.[8]

Because of such views, schools—the very institutions that should academically prepare youth for doing well in adult life—make little use of high-stakes examinations and effective incentives for accomplishments. School boards and administrators, for example, rarely measure and reward teachers' individual performance. Unions prevail in contracts that require paying public school teachers according to their degrees and years of experience, neither of which affects how much their students learn. After decades of declining union membership in other sectors, schools remain one of the few institutions that provide no merit incentives for their workforce.

The Social Promotion Disincentive

Examinations can allow educators to employ sticks as well as carrots. Consider the case of social promotion. Perhaps because the U.S. school system lacks accountability and incentives, students are usually promoted from one grade to the next whether they have or have not mastered the subject matter. Promoting failed students, however, does many harms. It wrongly informs them that they have learned what they need to know. It robs them of motivation. Why study if you know you will be promoted and graduate?

Such social promotion mixes failed students and successful students, which reduces teachers' effectiveness. They must teach things either that the successful students already know or that the

failed students are yet incapable of learning. In the long run, social promotion is unfair because the same high school diploma goes to all students whether they have earned it or not. This debases the value of diplomas—employers cannot depend on graduates' knowing what they should—and colleges and universities are forced to offer expensive programs and remedial courses for students to learn what they should have mastered in high school.

To reverse this common pattern, Chicago and other cities are instituting summer school for failing students. To be promoted, they must make satisfactory progress over the summer months. A longer school year seems appropriate in any case. In *A Nation at Risk,* the National Commission on Excellence in Education pointed out that the United States has the shortest school year in the industrialized world—only 180 days in contrast to about 200 in Europe and 240 in Japan.[9]

Harris Cooper of the University of Missouri, moreover, compiled thirty-nine studies of the "summer slump." The studies show achievement declines over the long summer vacation, especially among low-income, urban students.[10]

The first year of Chicago's summer program showed that the students gained substantially during the summer. Though about a fourth were not promoted to the next grade, they still had time to make up for lost ground.[11] They now also have the advantage of knowing that Chicago schools take standards seriously.

It is hardly a new insight that more study and classroom time increases learning. Indeed, it is one of the most consistent findings in educational and psychological research. Nearly all 130 surveys and experiments I compiled show the positive effect of more learning time.[12] What seems to be lacking are examination standards and incentives to elicit the learning time that would substantially raise achievement.

Defective Assessment Examinations

One of the latest harmful fads in education is "authentic" tests. Those who accept this terminology must subscribe to the view

that other tests are "inauthentic." As the term is often used in education circles, authentic tests consist of examinations that require recalled, or constructed, responses, as in essay questions, rather than examinations that offer a choice of correct answers among alternatives, as in standardized multiple-choice tests. Examples of authentic tests are oral examinations, laboratory exercises in science, musical and other performance exhibitions, and art and writing portfolios. Such so-called authentic tests, however, are hardly new; they have worked well in classrooms for decades if not centuries. What is new about them is using them in wide-scale surveys for school comparison, assessment, and accountability purposes.

Though the authentic-testing movement of the last decade is new and good, the new parts are not good, and the good parts are not new. The virtues of multiple-choice tests for large-scale assessment are that they are objective, reliable, valid, cheap, and hard to corrupt. They can widely sample students' knowledge of sixty ideas in as many minutes, whereas an essay examination may sample only one or two ideas. Multiple-choice tests can be made very difficult as in two- and three-step mathematics and science items. For these reasons, multiple-choice tests are most often employed in selection decisions for universities, graduate and professional schools, employment, and professional licensure in law, medicine, and other fields.[13]

In contrast, authentic tests used in large-scale assessments are easily compromised because a few essay questions or laboratory exercises are readily leaked. Even when students have mastered a few prespecified or leaked questions before examinations, they often do poorly on similar problems that are stated slightly differently or pertain to a different context. Also, in a given amount of time, such tests can only sample a limited number of ideas and skills. And finally, zealous parents can construct art, writing, and science portfolios.

These problems have long been known, and common sense would rule against the use of such examinations in large-scale assessments, particularly without small-scale trial assessments.

Nonetheless, it took very expensive, statewide trials of such examinations in California, Kentucky, and Vermont to prove what would seem obvious.

None of this is to say that traditional classroom examinations should be ruled out of large-scale assessments. But because they are far more expensive and rarely meet technical standards, they need to meet a high burden of proof of the additional value they bring to an assessment. Empirical studies of scores on well-designed "constructed-answer" tests show that they tend to rank students similarly to the rankings made by multiple-choice tests.

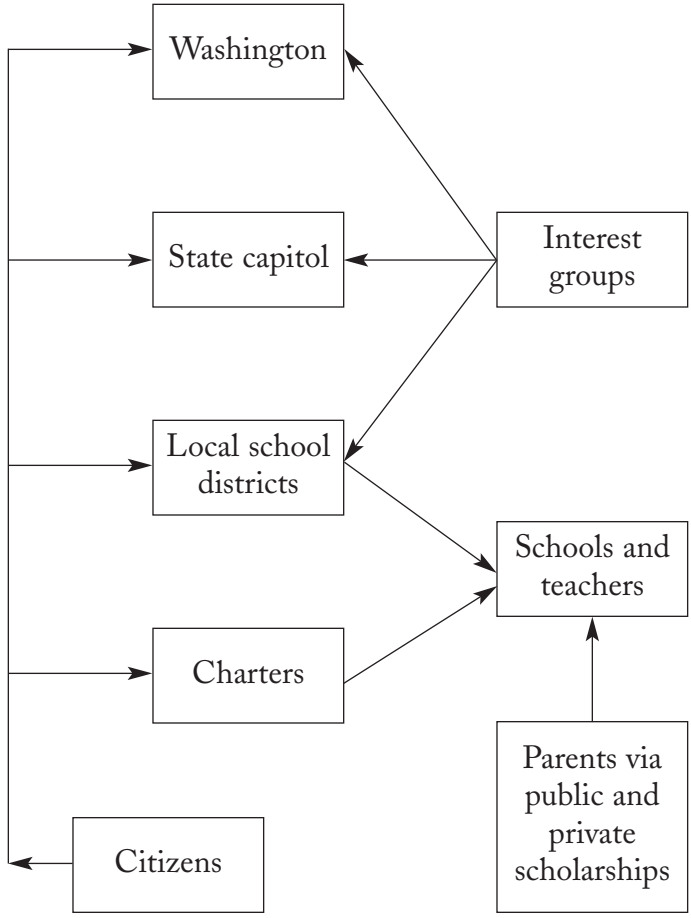
What "authentic tests" seem to bring to assessment, as demonstrated so far, is seemingly insuperable difficulties and prohibitively high costs. No wonder they are supported by those who wish to evade accountability.

The Governance Problem

The accompanying figure (Figure 1.1) broadly illustrates the governance problems that make U.S. schools unproductive. Ideally, schools should be as quickly and accurately responsive to individual consumer preferences as markets; they should reflect what local citizens and parents desire. But they can exert only indirect, cumbersome, slow-acting influence through elective processes of Congress, state legislators, and local school boards. What these governing agencies deliver might at best be an average preference of their constituents—but education tastes obviously vary.

Notwithstanding individual and local preferences, taxes are determined, collected, funneled, constrained, and filtered through national, state, and local governments. California, Hawaii, and the District of Columbia, for example, severely constrain variations on what local districts spend. States increasingly regulate local districts and induce them to follow what seems best in state capitols, which are often subject to special concentrated interests. So citizens cannot easily suit schools to their own preferences; their influence is indirect, attenuated, and interpreted by others.

FIGURE 1.1 Flow of Money and Influence



How state tax monies for schools are spent is strongly influenced by special interests, particularly teachers' unions, and the education administrators at the several levels. They have strong, well-defined interests, such as higher salaries, a short school year, an absence of competition, and no accountability—hardly identical with the public's interest or students' interests. Their interests make it worthwhile for status quo providers to follow

and influence legislation and policy implementation.[14] In contrast, parents cannot afford the time to study and modify voluminous, complex legislation and rule making that may affect their children.

The federal government pays only about 6 percent of the public school bill. It exerts much more influence, however, through regulations and various financial inducements for states to do things Washington's way. Federal categorical programs such as bilingual and special education grew mightily and created large self-interested bureaucratic units in state departments of education and local school districts, which is a major reason for the United States' having the highest administrative costs among affluent countries.[15] In any case, citizens' and parents' voices are hard to hear, and they are hardly a match for the status quo special interests.

How Can Examinations Help?

Three alternatives promise increased productivity: (1) employing more effective methods, (2) improving accountability, and (3) expanding local control, particularly through citizen and parent choice. Since choice enhances practices and increases accountability, it seems by far the best alternative.[16] Still, statewide public scholarships cannot be expected next month. Under the influence of special interests and the latest fads, Congress, state legislators, and school boards seem likely to increase accountability and regulate local practices by fiats, carrots, or sticks. Whatever combination of practices, accountability schemes, and choice ensues in a given city or state, achievement examinations can serve to evaluate results.

Examinations for Evaluating Education Methods

During the last half-century, scholars have published hundreds of randomized experimental studies of methods of instruction in dozens of journals, largely employing standardized objective examinations to evaluate the comparative results. During the past

ten years, much of this was reviewed in edited handbooks of research on instruction in science, mathematics, foreign language, and the other school subjects, some of them running to 900 double-columned pages by dozens of specialized chapter authors.

Synthesizing Research

A project sponsored by twenty-eight professional education organizations further condensed these works to 170 pages.[17] The specialized handbook editors, including my own work on generic methods, described in a page or two each of ten most effective teaching strategies. These include such traditional methods as direct instruction and mastery learning. Some newer methods, such as reciprocal teaching and employing "wait time" for students to answer science questions, are also discussed.

The evidence on the effect of the amount of time for study is particularly voluminous. Of 376 estimates of its effects, 88 percent were positive—one of the most consistent findings in education research. The sizes of the effects are moderate over a week or a semester; so time is hardly a short-term panacea. Over many years, however, student engagement in classes and in homework time yields huge benefits.

In a 1999 article,[18] I compiled the numerical sizes of effects of some 275 educational practices, some generic, others for separate subject matters. Some practices are several times more effective than others. Other things being equal, these should be chosen more often. Yet effective practices are often unemployed or poorly employed. Instead of relying on research, educators have often gone from fad to fad such as whole-language teaching, which had little basis in control-group research employing objective examinations.

Countering Biased Evaluations

Standard national examinations also can help with a growing problem of program self-evaluation. Some federally funded and foundation-funded groups have designed and evaluated their own programs. What would seem an obvious conflict of interest has been generally overlooked. Two of these programs, Success

for All and Reading Recovery, have reported superior results and have been widely popular despite their high costs. Their designers, however, carried out evaluations that in several ways biased the results in favor of their own programs.[19] Chief among these built-in biases was comparing their programs with others using tests that directly reflected what they were teaching. Combined with independent evaluations, standard examinations can help protect consumers in giving more objective estimates of program efficacy.

Identifying Promising Technologies

Although some were developed during the last decade, the effective programs discussed above are conventional and make little use of computer, electronic, and other technologies. Some technologies, however, offer the possibility of not only greater effectiveness but also lower costs and student convenience. We may immediately think of the Internet, but we have little solid research on its present or potential effects on learning. Distance learning may be a better example. As I learned in a survey I carried out for World Bank,[20] distance learning can free students from the limitations of space, time, and age and has a record of success in high- and low-income countries. Broadcast media, moreover, can multiply the effects of both books and traditional teaching.

Distance learning can include correspondence texts, books, newspaper supplements, posters, radio and television broadcasts, audio and video cassettes, films, computer-assisted learning, and self-instructional kits, as well as such local activities as supervision, supplementary teaching, tutoring, counseling, and student self-help groups. Scarce resources of scientific, pedagogical, and media expertise concentrated in development centers can be spread widely. The shortage of mathematics and science teachers in the United States and elsewhere are good reasons for employing distance programs.

Distance approaches can be highly cost-effective when large numbers of students follow the same preproduced courses. Far more than a single teacher working alone, distance courses can

incorporate validated subject matter and systematic instructional design and spread developmental costs over thousands of students. In rural areas such as Minnesota and Oklahoma, they can provide excellent courses in such subjects as calculus that would otherwise be unavailable. They can build on proven principles of individualized study by including clear learning objectives, self-assessment materials, student activities, and opportunity for feedback periodically or on demand. In high-density areas, the British Open University and the Chicago City Colleges have greatly enlarged opportunities for study, especially for those who cannot attend daytime classes.[21]

Objective examinations should prove helpful in identifying the best and most cost-effective of the old and new methods. Although cost, convenience, philosophical assumptions, and other considerations deserve weight, the consistency and magnitude of learning effects should be primary.

Examinations for Accountability

Since citizens pay for public schools, they should know how well the schools are doing. So also should parents, educators, legislators, and state and local school boards. To know this, they would need to compare the achievement scores of teachers, schools, and districts with one another and with standards of performance. James Coleman argued that if standards were externally set, as they are in many countries, educators could not lower them.[22] Teachers could then concentrate on helping their students meet the standards, as do coaches in competitive sports.

In addition, as Coleman emphasized, student heterogeneity needs to be taken into account if there are to be fair standards and accountability. Therefore, student progress or the value added to learning by the teacher or school during the most recent year or other time period should be the chief criterion.

The value added by a unit at each step can be the basis of accountability. This requires a measure at the end of each step so that the gains made by a student or group of students can be measured. Such value-added gains largely eliminate socioeconomic

and other extraneous differences and provide a fairer basis for evaluating progress.

As shown in Figure 1.1, the present cumbersome accountability system leaves the various governing units pressured by special interests without clear responsibilities. Complex regulations at the federal, state, local, and school levels not only slow decision-making but also remove the possibility of such clear accountability and responsibility. Education providers may benefit from the present system, but it might be in their long-term best interests as a profession to increase their productivity. Together with carrots, sticks, and choice, objective examinations, if well designed and used, seem likely to increase their accountability.

Examinations and Surveys for Evaluating Choice[23]

Objective examinations and other indicators of performance are proving useful in evaluating public and private scholarships, charter schools, and other means of choice. Various experiments in privatization and "contracting out" public services to private (including for-profit) firms suggest that they respond swiftly and accurately to citizens' desires. An economist's version of "meta-analysis" (analysis of results of many studies) shows that, other things being equal, private organizations on average perform better at lower costs and that they are more satisfying to their staff and their customers.[24] Requiring performance, satisfaction, and cost indicators, these studies concern airlines, banks, bus services, debt collection, electric utilities, forestry, hospitals, housing, insurance sales and processing, railroads, refuse collection, savings and loans, slaughterhouses, water utilities, and weather forecasting. In the United States and other countries, governments are privatizing prisons, police and fire protection, and public pensions.

Objective examinations and other indicators already suggest that choice works similarly in education. One instance in the United States of a randomized experiment employing publicly funded scholarships for school students is Milwaukee.[25] Minority-children applicants to private schools were admitted at

random because there were insufficient spaces to meet demand (even though the costs of the program were about half that of public schools, the state department of education created many uncertainties and difficulties, and various start-up difficulties ensued).

After several years, the selected students did significantly better on standardized tests, sufficiently well that the usual national minority-majority achievement gap could be substantially cut. Perhaps even more important and notwithstanding the start-up difficulties, parents were delighted with the private schools.

Private schools, moreover, improve public education because what public schools lack is competition; most are local monopolies. Recent research supports the value of local private competition in improving public schools. Rather than "creaming off" the best students, the presence of private schools and public schools of choice is associated with better examination performance and lower costs among nearby public schools.

Examinations and Consumer Preferences

Surveys can usefully supplement objective examinations in evaluating choice. They show that citizens increasingly favor choice in education, including private schools. A 1992 Gallup poll showed that of those polled, 70 percent supported publicly funded scholarships that include private school choice. Eighty-five percent of African Americans and 84 percent of Hispanics supported such scholarships.[26] Big-city poor and minorities particularly favor scholarships because government schools available to them are most often dominated by teachers' unions, federal categorical programs, and regulations, which make their schools unresponsive to their preferences. In Milwaukee, for example, which has had the most experience with privatization, 95 percent of African Americans favored private and public school choice, and 70 percent believed that students get a better education from independent and sectarian schools.[27]

This accords well with my recent experience as a board member of the Chicago Charter School Foundation, a publicly funded and privately governed and operated charter school of about

4,700 mostly African American and Hispanic students on seven campuses. Unlike most charters in Chicago and elsewhere, the not-for-profit board contracts with for-profit and other providers[28] of the education program. Despite the usual start-up difficulties of finding, purchasing, and refurbishing buildings and hiring an entirely new staff, the campuses have long waiting lists. Our students comprised 70 percent of the students in Chicago's charter schools. They scored best on standardized examinations among students in the charter schools.

Other behavioral indications of private preference can be cited: Consider the rise in home-schooled children from 300,000 in 1990 to about 1.5 million today. Their parents give attention, time, and money for what they think is a superior education.[29] Though they are mostly amateur teachers, their children's average achievement on standardized tests exceeds 77 percent of regular school children.

Another indication of preference for private over public education is the quiet but meteoric growth of privately funded scholarships, often concentrated on poor, big-city children. Funded by individuals, philanthropies, and firms, their numbers in seventeen cities have grown from 744 to 6,572 in a recent four-year period—further evidence of consumer preference that complements what we can learn from examination performance.[30]

Reconciling Choice and Standards

Those who favor choice should know that most American schools are unlikely to be funded through public and private scholarships today. Powerful special interests—the teachers' unions and other education lobbying groups—adamantly oppose school choice, especially free choice by parents. These groups strongly influence legislators, school boards, and the mass media. They have, for example, severely restricted the number and nature of charter schools.

For choice advocates, it seems reasonable to avoid both incrementalism and radicalism while collecting more evidence. Aside from consumer preference, the best evidentiary criteria are high scores on objective achievement tests in the standard school subjects. These should be employed in experiments to evaluate the many variations on choice that have been proposed. Once these

experiments have been completed, there may be less need for evidence. Even so, it seems likely that policy makers and parents would continue to want to know about how well students perform on objective tests.

Vast amounts of research and common experience suggest that markets and competition work well; they increase effectiveness, lower costs, and better satisfy citizens who are free to choose. Providers of public education insist, nonetheless, that education is an exception. They claim to protect the poor and minorities even though these groups most strongly favor and benefit from choice. They oppose trials of choice but insist on more evidence before it is tried.

As citizens of a country that has the near highest per-student costs and the worst value-added achievement gains,[31] Americans have the right to ask where the burden of proof lies. Objective examinations, parent satisfaction surveys, waiting lists for places, and other indicators have important roles in better answering the questions before us.

Notes

1. Herbert J. Walberg, *Spending More While Learning Less* (Washington, D.C.: Thomas B. Fordham Foundation, July 1998).
12. Organisation for Economic and Co-operative Development, "Education at a Glance 1995" (Paris: OECD, 1995).
3. For further examples, details, and documentation for this section, see my "Uncompetitive American Schools: Causes and Cures" in ed. Diane Ravitch, *Brookings Papers on Education Policy* (Washington, D.C.: Brookings Institution, 1998).
4. Herbert J. Walberg and Herbert J. Walberg III, "Losing Local Control: Is Bigger Better?" *Educational Researcher* (June/July 1994) 23–29.
5. Ibid.
6. Caroline Minter Hoxby, "How Teachers' Unions Affect Education Production," *Quarterly Journal of Economics* (August 1996): 1–24; Sam Peltzman, "Political Economy of Public Education: Non-College-Bound Students," *Journal of Law and Economics* 39 (1996): 73–120.
7. For Public Agenda surveys of teacher educators and others cited in this section, visit *http://www.publicagenda.org/aboutpa/aboutpa2p.htm*.

8. Nearly all public school teachers are paid according to their degrees and experience—neither of which influences their students' achievement—rather than academic mastery, the use of effective practices, their students' achievement, or other indicators of merit.
9. National Commission for Excellence in Education, A *Nation at Risk: The Imperative for School Reform* (Washington, D.C.: U.S. Government Printing Office, 1983).
10. Harris Cooper, *Making the Most of Summer School,* Monograph Series of the Society for Research in Child Development (Malden, Mass.: Blackwell, 2000).
11. Julian R. Betts and Robert M. Costrell, "Incentives and Equity Under Standards-Based Reform," in *Brookings Papers on Education Policy 2001*, ed. Diane Ravitch (Washington, D.C.: Brookings Institution, 2001), 9–74.
12. Herbert J. Walberg, "Uncompetitive American Schools: Causes and Cures" in *Brookings Papers on Education Policy*, ed. Diane Ravitch (Washington, D.C.: Brookings Institution, 1998), 173–206.
13. For more detailed information on standardized and traditional tests and on assessment reform issues, see Herbert J. Walberg, Geneva D. Haertel, and Suzanne Gerlach-Downie, *Assessment Reform: Challenges and Opportunities* (Bloomington, Ind.: Phi Delta Kappa, 1994).
14. Myron Lieberman finds the AFT and the NEA are among the most powerful and sophisticated interest groups. They enroll more than three million members whose dues exceed $1 billion annually. Lieberman calculates that they employ more political operatives than the Democratic and Republican parties combined. Their 405 representatives at the 1996 Democratic convention exceeded all states but California. See *The Teacher Unions* (New York: Free Press, 1997), 2–5.
15. Block grants of money for education returned to states might simplify educational administration; such grants would probably reduce costs and special interest influence and increase state and local accountability, control, and productivity. Another administratively simplifying alternative is assigning goal setting and monitoring to states and assigning operations to local districts or schools. In addition, charter schools in principle strengthen direct, local control, but many appear hampered by regulations, union contracts, and initial capital costs. For a variety of reforms, including national standards, charter and for-profit schools, vouchers, New American Schools, and the especially interesting cases of Charlotte-Mecklenberg, Chicago, San Diego, California, Kentucky, and Minnesota, see Chester E. Finn Jr. and Herbert J. Walberg, eds., *Radical Education Reforms* (Berkeley, Calif.: McCutchan Publishing, 1994).

16. See Herbert J. Walberg and Joseph Bast, "School Choice: The Essential Reform," *Cato Journal* (spring/summer 1993): 101–22.
17. Gordon Cawelti, ed., *Handbook of Research on Improving Student Achievement* (Arlington, Va.: Educational Research Service, 1995).
18. Herbert J. Walberg and Jin-Shei Lai, "Meta-Analytic Effects for Policy" in *Handbook of Educational Policy,* ed. Gregory Cizek (San Diego, Calif.: Academic Press, 1999).
19. Herbert J. Walberg and Rebecca C. Greenberg, "The Diogenes Factor: Why It's Hard to Get an Unbiased View of Programs Like 'Success for All,' " *Education Week* 52: 36.
20. Herbert J. Walberg, "Improving School Science in Advanced and Developing Countries," *Review of Educational Research* 61 (1991): 25–69.
21. Because of some causal uncertainties, a few early negative studies, and uncertain cost figures, it would seem advisable to conduct a comprehensive review of all accessible distance-education studies. Such a review would not only estimate overall comparative effects and costs under varying conditions, but also identify the most effective practices and combinations of media.
22. James S. Coleman, "Achievement Oriented School Design," paper prepared for the Social Organization of Schools Conference held at the Center for Continuing Education at the University of Notre Dame, March 19, 1994.
23. For further examples, details, and documentation for this section, see my "Uncompetitive American Schools: Causes and Cures" in *Brookings Papers on Education Policy,* ed. Diane Ravitch (Washington, D.C.: Brookings Institution, 1998).
24. Charles Wolf, *Markets or Governments: Choosing Between Imperfect Alternatives* (Cambridge, Mass.: MIT Press, 1988), 137–48.
25. Paul Peterson of Harvard University has been analyzing parent satisfaction and examination performance of students given or not given vouchers by lottery in Dayton, Ohio, New York City, and Washington, D.C.
26. Stanley M. Elam, Lowell C. Rose, and A. M. Gallup, *The 24th Annual Gallup/Phi Delta Kappa Poll of the Public's Attitudes Toward the Public Schools* (Bloomington, Ind.: Phi Delta Kappa, 1992).
27. Nina Shokraii, "Free at Last: Black America Signs up for School Choice," *Policy Review* (December 1996):20–26.
28. In my view, for-profit firms offer exciting prospects for improving productivity. They must furnish entrepreneurial talents and energies to satisfy their customers. Unlike public schools, if they fail to please, they go out of business. If they please

parents (and, we could add, meet achievement goals), they deserve their profit in return for invested risk and the ability to serve. If subject to competitive bidding, they would probably reduce costs while improving achievement, satisfaction, and other outcomes.

29. Barbara Kantrowitz and Pat Wingert, "Learning at Home," *Newsweek* (October 5, 1998): 64–71.
30. Terry M. Moe, *Private Vouchers* (Stanford, Calif.: Hoover Institution Press, 1995), 14.
31. Herbert J. Walberg, *Spending More While Learning Less* (Washington, D.C.: Thomas B. Fordham Foundation, July 1998).

Chapter 2

Why Testing Experts Hate Testing

Richard P. Phelps

Introduction

The public has often been asked how it feels about testing. Over several decades and in a variety of contexts, the American people have consistently advocated greater use of standardized student testing, preferably with consequences for failure (that is, high stakes). The margins in favor have typically been huge, on the

The author would like to thank: the Fordham Foundation for its support; Chester E. Finn Jr., Marci Kanstoroom, Diane Ravitch, Steve Ferrara, and Mike Petrilli for their very helpful edits; Scott Oppler, Chris Sager, and Deb Wetzel for advice on certain psychometric concepts; and James Causby, Karen Davis, Kathleen Kennedy Manzo, Vanessa Jeter, and Janet Byrd for supplying important information about North Carolina's testing program. The author retains all responsibility for errors.

This chapter originally appeared as *Fordham Report* 3, no. 1 (January 1999). It has been slightly revised and updated.

order of 70-point spreads between the percentage in favor of more testing and the percentage against.[1]

But the public may not get its way. Many educators and education "experts" oppose standardized testing and high stakes. Although this throng includes some school administrators who fear the fallout from poor test results, it also, and most notably, includes many education school faculty members. In a 1997 survey, a national sample of them voiced substantially less support for high-stakes standardized testing than did other groups. "[O]nly 49 percent believe raising the standards of promotion from grade school to junior high and letting kids move ahead only when they pass a test showing they've reached those standards would do a great deal to improve academic achievement. In sharp contrast, the percentage reaches 70 percent among the general public [and 62 percent among teachers]."[2]

The polling organization Public Agenda found that "while supporting standards in concept, professors of education seem reluctant to put into place concrete, high-stakes tests that would signal when kids are meeting the standards."[3] They are especially opposed to multiple-choice tests. "Fully 78 percent want less reliance on multiple-choice exams in the schools. . . . [E]ducation professors . . . call for more reliance on portfolios and other authentic assessments."[4]

These faculty members don't think standardized tests demonstrate learning. "The fact is that all of the data say standardized tests don't predict what they are intended to. They just don't do it. . . . There is no standardized test that is good," a Boston professor told Public Agenda.[5] The professors recognize that the public has a different view of testing, however. Public Agenda reported that many faculty members expressed "disappointment and some exasperation that so much current educational research seems to be ignored or dismissed by the public."[6]

In June 1991, the American Educational Research Association (AERA), a group consisting primarily of education professors, hosted a press conference on student testing issues in Washington, D.C., as a "public service to build bridges between researchers and policy makers." Five prominent members of the

group presented papers, such as "The Teacher, Standardized Testing, and Prospects of Revolution," in unanimous opposition to President George H. W. Bush's then-pending proposal for national tests; high-stakes use of standardized tests; multiple-choice formats; "external" tests (that is, tests not wholly controlled by school staff); and other features of student testing that they disliked.[7]

The reader may be struck by a paradox: It frequently seems that experts on testing have never met an actual test that they like and want to see used. What is it about testing that troubles them so? At the AERA press conference, a now-familiar litany of assertions was offered to explain how "research" shows that standardized testing is bad. The antitesting canon includes allegations that standardized tests, particularly those with high stakes,

- induce "teaching to the test," which, in turn, leads to artificial inflation of scores.
- narrow the curriculum to a small domain of topics.
- tap only "lower-order thinking" and hence discourage innovative curricula and teaching strategies.
- cause student achievement to decline.
- are unfair to minorities and women.
- are costly in terms of money and time.
- are overused in the United States, especially in comparison with other countries.
- are opposed by all who truly care about children.

Not all testing experts dislike testing, however. Hundreds of them work cheerfully for state and local testing agencies and for test developers. The opponents we hear from the most are a relatively small group of "testing policy" researchers, who are on the faculty of education schools or who work at organizations such as the federally funded Center for Research on Evaluation, Standards, and Student Testing (CRESST),[8] the Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP)[9]

at Boston College, and an advocacy group known as the National Center for Fair and Open Testing (FairTest).[10] A brief excerpt from a FairTest publication entitled *Fallout from the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools* sums up the basic position of the organization.[11]

> Standardized tests often produce results that are inaccurate, inconsistent, and biased against minority, female, and low-income students. Such tests shift control and authority into the hands of the unregulated testing industry and can undermine school achievement by narrowing the curriculum, frustrating teachers, and driving students out of school. [12]

In this chapter, the arguments of the testing experts who dislike testing are held up for careful scrutiny. First, four case studies that suggest how these experts deploy their arguments in the real world are examined. Then those arguments are appraised.

Case Study 1: The National Assessment of Educational Progress

The nominal reason for the AERA press conference was to criticize then-President George H. W. Bush's national testing proposal. Proposals for national testing systems, be they from George H. W. Bush or Bill Clinton (George W. Bush has not proposed any additional national test), tend to attract a great deal of attention. To date, however, there has been only one such test—the National Assessment of Educational Progress (NAEP). The NAEP is an assessment based on samples of schools, and no individual student information is made available. It is a no-stakes test.

For decades, NAEP samples were exclusively national and so were NAEP scores. In the 1980s, however, many people pressed for state-representative NAEP samples ("State NAEP"). Almost half the states had instituted their own testing programs, many of them high-stakes "minimum competency" graduation requirements. Some state leaders wanted to gauge their students' levels of achievement or the progress of their states' education reform efforts against an external benchmark, and the scores of state-

representative samples of schools and students on the NAEP seemed the perfect candidate to be that benchmark.

But what sounds like an obvious idea drew strong opposition from testing experts. Daniel Koretz, a researcher with CRESST and the RAND Corporation, made three separate arguments against releasing state-by-state NAEP scores. First, he argued that the public cannot be trusted with such information. Koretz wrote that "[S]ome differences among states would be too fragile—too dependent on the specifics of the test—to warrant the simple interpretations that they will receive."[13] Second, Koretz argued, academic success is predicted primarily by the socioeconomic background of students, so state-level NAEP will just show once again that richer states do better:[14] "To infer that a difference between two states on the NAEP reflects specific policies or practices, one needs to be able to reject with reasonable confidence other plausible explanations, such as economic or demographic difference."[15] (Other opponents of State NAEP have made these arguments even more forcefully.[16]) Third, Koretz insisted that because State NAEP provides only cross-sectional data, it cannot show improvements in achievement that may coincide with education reform programs: "NAEP is purely cross-sectional, which eliminates a large number of the designs that could be used to draw causal inferences."[17]

The essence of these objections is that state-level NAEP results would be used to judge states and these judgments would inevitably be unfair. So, because people who don't understand what the scores really mean would use this information to evaluate the states, we shouldn't gather the information at all.

In a counter to Koretz, Gary W. Phillips of the U.S. Education Department's National Center for Education Statistics noted that although a single administration of State NAEP might not allow us to evaluate the impact of reforms, a system needed to be established that could be used to appraise such changes in the future. We had to start somewhere.[18] The National Academy of Education, assigned to review the efficacy of State NAEP, recommended implementation and reiterated that recommendation in a 1996 review.[19]

With several administrations of state-level NAEP now behind us, we have time-series data with which to gauge the progress (or lack thereof) that each state's youngsters are making in mathematics, reading, and science. We can thus begin to see where state education policies are effective, with background factors controlled. The utility of NAEP scores as markers for monitoring state education reforms is seen in the next two case studies, of Texas and North Carolina.

The 1988 legislation establishing state-level NAEP also permitted "standards-based" reporting of scores. Historically, NAEP results were reported only according to abstract "scale scores" that were not anchored to any standards. But the National Assessment Governing Board—to the continuing dismay of many testing experts—judged that NAEP results would be far more useful, particularly in tracking progress toward the national education goals that the President and governors set in 1989, if they showed how U.S. children were doing academically in relation to how well they *ought* to be doing. The Governing Board established three performance levels, which it termed "basic, proficient, and advanced," and accompanied each with descriptions written in English about the specific skills and abilities represented by each level. Like State NAEP, the performance level concept and the method for setting the levels have drawn controversy, with some testing critics favoring the old scale scores' aloof abstractness and many policy makers desiring more useful and understandable measures.[20]

It would appear, however, that the performance levels are here to stay. The National Assessment Governing Board has remained steadfast. And a National Research Council review of NAEP, while agreeing with the critics on a number of specific points, also conceded: "It is clear that Americans want the kind of information about the achievement of the nation's students currently provided by NAEP summary scores and achievement-level results."[21]

Case Study 2: Texas

Perhaps no state testing program has aroused the ire of testing critics more than the Texas Assessment of Academic Skills

(TAAS), since 1990 the backbone of the Lone Star State's education accountability program. In its ratings of all state testing programs, FairTest rated the TAAS at 2 on a scale of 1 to 5, with 1 being the worst score possible.

FairTest explained its dim view of TAAS as follows:

> The Texas assessment system needs many major changes. It relies almost entirely on multiple-choice items, except for a writing prompt, and has a high-stakes graduation test. On most of the other standards, however, the state does very well. It has strong bias review procedures, provides solid public information, accords parents substantial rights, and has a thorough and continuing review system. Professional development appears fairly extensive.[22]

Observe that FairTest gave the state's testing program the second-lowest possible rating for only two reasons: high-stakes and multiple-choice formats. According to Monte Neill of FairTest, "When you have high stakes and then add an exit exam, that jacks up the system so that the test becomes the curriculum. . . . One should not be using scores on tests to make serious educational decisions."[23]

Responding to evidence that pupil achievement in Texas has improved markedly since the TAAS was introduced, Neill "concedes that the improvements are impressive," reports *Education Week,* "but he says that an enriched curriculum, not test preparation, is behind the shifts."[24] There may be a contradiction here. According to Neill, the test has *become* the curriculum in Texas and the improvement in student achievement is the result of an enriched curriculum. Still, he declines to see the improvement as linked to the testing.

In addition to FairTest's criticisms of the TAAS for its high-stakes and multiple-choice formats, the Texas testing program was the subject of two separate lawsuits. The NAACP asserted that it was biased against blacks since they performed worse than whites on the test.[25] The Mexican-American Legal Defense Fund followed with a suit using the same logic.[26] Both cases were heard by the U.S. Education Department's Office for Civil Rights and Texas state courts and were dismissed.[27]

Through the clouds of flack, the citizens of Texas remained on course, retaining and expanding the TAAS. Moreover, the results

do appear to be positive. Texas students' average state test scores have shown achievement gains year after year. That Texas students have also made gains well above the national average on the NAEP throughout the past decade would seem to corroborate the improvement.[28]

Other benefits have also followed. Observers of Texas education report

- a greater focus on academic learning.
- a culture of high expectations and enthusiasm toward reaching standards.
- generous and immediate remediation efforts offered to poorly performing students, both because a system is in place to identify their problems early and because, with high stakes, students' problems are not just passed along to the next grade, where they become compounded.
- greater interest among teachers in academic strategies and more cooperation with each other to learn which ones work best, and how.
- that with a regular system of assessment, school staff can get quicker feedback on which instructional systems work best.
- Texas has built a school-specific information system on the World Wide Web for all parents to see, helping them understand their schools better.[29]

Though always intended to match Texas's curriculum and performance standards, the state's student testing program first took aim at basic skills and minimum competency, a focus that may have neglected the more advanced students. It is now being expanded to cover more grades and purposes (statewide high-school-level end-of-course examinations, for example). It has strived to achieve better integration with the curriculum, professional development, and program planning, as well as with student evaluation, and is today a key component of one of the most comprehensive accountability systems in the country.[30]

Texas's accountability system has received strong political support from both parties. Republican and Democratic governors alike have resisted most attempts to soften its requirements, even in the face of sustained criticism. Indeed, gubernatorial opponents in the 1994 election attempted to outdo each other in their support for still higher standards and tougher requirements.[31] In 1998, it was not even an issue.

Case Study 3: North Carolina

A similar story can be told about North Carolina, a state that, like Texas, ranked near the bottom on the NAEP but has improved its student achievement dramatically after instituting a comprehensive, integrated, high-stakes testing program and sticking with it despite serious opposition.[32]

The North Carolina Education Department rates schools based on their results on state tests. It is a value-added rating system in which adjustments are made—for socioeconomic and other background factors—to the expected performance of each school. Teachers at schools rated "exemplary" are rewarded monetarily. But poorly performing schools are not abandoned. The department assembles teams of three to five experts in curriculum and instruction who work with those schools for an entire year. These teams help school staff align their curricula with state academic standards, and they also demonstrate effective teaching techniques and try to locate additional resources for the schools.[33] Fifteen schools designated for "mandatory assistance" at the end of the 1996–1997 school year finished 1997–1998 by achieving "exemplary" ratings for improving their performance by more than 10 percent.[34] In 1998–1999, state "assistance teams" visited forty-six public and seven charter schools, eleven of them under a "mandatory assistance" provision for the worst-performing schools in the state, the rest under voluntary arrangements.[35]

The sixty-odd schools visited by state assistance teams represent less than 3 percent of the state's schools. Most schools either develop their own programs or rely on assistance from their school district. Jeff Moss, the associate school superintendent in

Hoke County, one of the state's poorest districts, describes how his district treats low-performing students there:

> . . . on a Friday the students take a test. If they do not pass the test, they have to come back after school hours the following week, Tuesday, Wednesday, and Thursday. They take another test the next Friday. They'll get the higher of the two grades; we don't penalize them for coming back after school hours.[36]

Students even get a third chance with the help of another round of after-school classes. A Southern Regional Education Board study of the Hoke County schools found that:

> . . . the percentage of students who now meet the state's algebra proficiency standard has doubled. Twenty percent more now meet the history standard. And the high school's overall Scholastic Assessment Test (SAT) scores are up 11 percent over three years. Also . . . employers are more welcoming of graduates now.[37]

The whole process of reform in Hoke County was set in motion by its initial poor showing in the state testing program, which identified the district's academic problems.

Still, holding students, teachers, or schools to fixed standards means that some will do less well than others and may be held back. Johnson County, North Carolina, for example, passed a student-accountability policy of its own in 1996. The policy called for intensive remediation, but also for retention in grade of students who did not score at a proficient level on state exams.[38]

District officials claim the accountability program has boosted student performance. According to Johnson County officials, more than one-third of students performed below grade level on the tests four years ago, yet just 1 percent of students were held back.[39] This year, less than one-fourth of the students performed below grade level and 8.8 percent of students were retained under the policy—and for other reasons, such as absenteeism. The other 16 percent were promoted "based on the grades they earned and other academic factors."[40]

Not everyone liked the new policy. Fourteen parents filed suit against Johnson County on behalf of children who were held back. They argued that the tests were intended by the state to rate

districts and schools, not individual students, and thus were "not valid for measuring individual performance."[41]

Walter Haney, a researcher with CSTEEP at Boston College, agreed:

> It is a prime example of a test that was developed for one purpose . . . and applied for a purpose that is totally inappropriate and unintended. . . . The North Carolina end-of-grade tests were designed to hold schools and districts accountable. There is considerable potential for people trying to use [a] national test for similar decisions without stopping to examine whether, in fact, the content parallels the local curriculum.[42]

The North Carolina tests do match state curriculum standards, however, and cover a representative sample of it. Because the state uses the tests to evaluate districts and schools, individual students usually see only one-third of each subject-area exam; by sampling this way, the state can cut testing time and costs. Had Johnson County held students back for poor performance on a test that covered only one-third of the curriculum, that would have been unfair. Instead, the district put the three separate pieces of the exam together to form complete exams that covered the entire curriculum.[43]

A U.S. District Court judge rejected the plaintiffs' request for an injunction to prevent another year of student retention.[44] The plaintiffs later dropped the case.[45]

Richard Jaeger, one of the speakers at the AERA press conference mentioned earlier, who was then a professor at the University of North Carolina, criticized his state's testing program in general, but particularly its high-stakes, minimum-competency, elements; the testing program is geared toward a relatively low level of basic skills and students have several chances to pass, starting in tenth grade. They cannot graduate from high school until they pass it.

Jaeger argued that the costs to society of denying students diplomas might be too high. "As a determinant of a student's life chances in American society, possessing a high school diploma is far more important than scoring well on a basic skills competency test." [46] He cited statistics showing that high school dropouts are

more likely to have blighted lives and argued that "the use of such tests jeopardizes the future of those young people denied a high school diploma by limiting their employability, reducing their quality of life, and diminishing their opportunity to contribute to society through the productive applications of their abilities."[47] Jaeger also presented evidence purporting to show that meeting higher standards and passing high-stakes tests do not improve students' economic prospects. He implied that if North Carolina just gave poorly performing students their diplomas with no impediments, the state would enjoy less crime, fewer out-of-wedlock births, and shorter welfare rolls.[48]

Two other presenters at the AERA conference also accused high-stakes tests of increasing the dropout rate.[49] Their evidence, however, was spotty. Most U.S. dropouts leave school when they reach the limit of the compulsory attendance law, not when they fail an exam.[50] When students in the large-scale Indiana Youth Poll explained why some dropped out, either disinterest in school or non-academic-related problems (such as pregnancy or family problems) were cited more than four times more often than academic failure.[51]

A careful examination of the dropout issue by Bryan Griffin and Mark Heidorn, using data from Florida from the early 1990s (when a test similar to the one used in North Carolina was in place), found that:

> . . . failure on a [minimum competency test] provided a statistically significant increase in the likelihood of leaving school, but only for students who were doing well academically. Students with poorer academic records did not appear to be affected by MCT [minimum competency test] failure; similarly, minority students did not demonstrate an increased likelihood of leaving school as a result of failing an MCT.[52]

More recent studies of the relationship between high-stakes tests and the dropout rate have shown that it can move in either direction—the dropout rate can rise if struggling students are ignored or decline if they are given the attention they need when they need it.

Speaking about the same high-stakes exit exam in Florida, psychologist and attorney Barbara Lerner explained:

> On the first few tries, 80 to 90 percent of Florida's students failed the test. But they were not crushed, as the experts predicted, and they did not give up and drop out in droves without diplomas. They kept trying, and their teachers did, too, working hard to help them learn from failure and, ultimately, to master the skills they needed to graduate. By the fifth try, better than 90 percent of them did just that. They left school not just with a piece of paper, but with basic skills that prepared them better for life.[53]

In spite of great advances in student achievement linked to the testing system in North Carolina, however, FairTest gave the state's system its lowest rating of 1 (on a scale of 1 to 5):

> North Carolina's assessment program needs a complete overhaul. It relies far too heavily on multiple-choice tests, tests too often, and has a graduation exam. It should reduce the grades tested, drop the graduation requirement, ensure districts do not rely on the tests for grade promotion decisions, and implement a performance assessment system based on the state standards.[54]

Overall, North Carolina showed the most improvement of any state on the NAEP in the 1990s.[55]

Case Study 4: SAT I (formerly the Scholastic Assessment Test)

The test attracting the loudest and most sustained opprobrium from critics over the years is the SAT I (formerly the Scholastic Assessment Test), used by almost two-thirds of U.S. colleges in making admissions decisions.[56]

One of the primary sustaining causes of FairTest is its crusade to convince colleges to cease using SAT scores in admissions decisions. If one read only FairTest's literature, one might well conclude that the group's campaign against the SAT has been very successful.[57] According to FairTest, hundreds of colleges now have optional or limited SAT requirements.[58]

Those colleges that offer the possibility of admissions sans test scores may, however, require additional proof of ability, such as a

graded writing sample or an on-campus interview. Moreover, even if not *required* for admission, the absence of a test score may still bias an application negatively.[59]

Still, the SAT's impact is often overstated. The overwhelming majority of colleges are not selective, so a low SAT score will rarely keep a student out of college. Even at the most selective colleges, the SAT is seldom used alone by college admissions staff to make decisions. Typically, it is one of many factors, which include a student's high school grade point average, extracurricular activities, recommendations, essays, and so on.[60] When surveyed, however, admission counselors rate the SAT score as a more reliable measure than these other indicators.[61]

The primary argument of SAT critics pertains to the test's "predictive validity"; it explains only 15 percent of the variation in first-year college grades, after other predictive factors are accounted for.[62] If that's all the good it does, why bother with it? they ask.[63] As Haney of CSTEEP says:

> Which is more accurate? Does a person's height more accurately predict a person's weight? Or do national college entrance exams more accurately predict a student's success in college?
>
> The answer: Height is a better predictor of weight. And there might be some crude relationship between height and weight. But it ain't real good.[64]

To a college admissions counselor, however, 15 percent is a lot of predictive power, and the SAT costs only about $20.65 It costs society about $25,000 to educate a high school student. For an incremental cost of 0.08 percent over the cost of a high school education, the SAT score provides a college admissions counselor a 34 percent increase in information. The incremental benefit-cost ratio for the SAT is 425:1 over the high school record.[66] The break-even value of the SAT is more than $8,500 per student; at $20, it's a bargain.

The SAT is a nationally *standardized* measure; a grade point average is not. One student can achieve a high grade point average by working extremely hard in difficult courses in a high school with exacting standards, while another can get by choosing easy courses at a high school with low standards.

Ultimately, the makers of the SAT do not determine its success; its customers do. Those customers are thousands of college admissions officers throughout the United States who are doing their best to select students they believe can handle the level of academic rigor at their institution.

College admissions officers are not ignorant. They hear and read the arguments against use of the SAT. Nor are they elitist conspirators opposed to fair admission policies. Moreover, they are not required to use the SAT (or the ACT). They use such tests because they believe, based on personal experience, that they are valuable—so valuable that they consider test scores to be the second most important criterion in making admissions decisions, higher than grade point averages or class ranks, and second only to grades and test scores from Advanced Placement courses, the only other nationally standardized measure of achievement commonly available to them.[67]

Appraising the Criticisms

The basic argument made by testing critics is that high-stakes standardized tests are counterproductive. Instead of leading to stronger academic achievement, they actually interfere with good teaching and learning. Testing experts embrace a sort of domino theory: Pressure to produce higher scores leads teachers to focus on material that will be covered by the tests and to exclude everything else.[68] The curriculum is thereby narrowed, which means that some subjects are ignored. Within those that are taught, lower-order thinking skills are emphasized because these are what the tests tap. As a result of teachers teaching to the tests, subsequent test scores are inflated while real learning suffers.

In addition to the alleged harms of (1) test score inflation, (2) curriculum narrowing, (3) emphasis on lower-order thinking, and (4) declining achievement, testing experts add a quartet of other arguments against testing: (5) that standardized tests hurt minorities and women, (6) that the tests are too costly, (7) that other countries don't test nearly as much as we do, and (8) that parents, teachers, and students in this country are all opposed to

testing. These eight claims are examined in detail in the section that follows and a rebuttal is offered to each.

What testing experts do not like are high-stakes, multiple-choice, external tests. These tests are excoriated with bad-sounding words ("lower-order thinking," "factory model of education," "uncreative," "rote recall," and so on), but such terms are seldom well explained. The root of most objections to testing can be traced to the dominant worldview of testing experts (and many other educators).

The education philosophy driving many of these criticisms is constructivism, the view that every student and teacher constructs his or her own meanings from classroom activities, books, and so on. Hence no construction is wrong or bad. We all know that there is often more than one way to get to a right answer. We all think differently, using different combinations of several different kinds of intelligence. Moreover, we all know that a student can process much of a problem well but still get the "wrong" answer in the end because of a fairly minor error, such as misplacing a decimal point.

As test critic (and constructivist) Mary Lee Smith of CRESST and Arizona State University describes it:

> Constructivist theory assumes that students construct their own knowledge (rather than passively receiving knowledge transmitted by school) out of intentional transactions with materials, teachers, and other pupils. Learning is more likely to happen when students can choose and become actively engaged in the tasks and materials and when they can make their own connections across subject matter on tasks that are authentic and organized around themes. According to this theory, literacy is whole, embodying reading authentic texts and writing as a way of unifying all the subjects. For example, to be literate is to be able to explain the reasoning one uses to discover and solve math problems. Explicit in constructivist theory is the rejection of the pedagogy of worksheets and the exclusive reliance on phonics, spelling out of context, computation, isolated subject matter and the like.[69]

Constructivists oppose school practices that they think "fix" behavior. They see standardizing curricula and instructional prac-

tice as restricting teacher behavior and multiple-choice standardized tests as shackling student responses to problems.

For constructivists, the more open-ended the assessment the better, and portfolios are the most open-ended of all. They involve no standardized, mandated preset response and not necessarily even a standardized question to impede any student's unique understanding of the problem, creative solution, and personal construction of the work.[70] This constructivist worldview is seen to underlie most of the arguments marshaled by testing experts against testing.[71]

1. Test Score Inflation

An initial set of harms ascribed to standardized testing fall under the rubric of teaching to the test. A CRESST paper entitled "The Effects of High-Stakes Testing on Achievement," by Daniel Koretz, Lorrie Shepard, and others, purports to demonstrate that high-stakes tests in fact cause teaching to the test.[72] The researchers compared student performance in math and reading from one commercial test given under high-stakes conditions with student performance on a different commercial test with no stakes. Student performance on the high-stakes test improved over time, according to the researchers, as the teachers adapted their instruction to the curriculum implicit in the test. Student performance on the other test, administered solely for the purpose of the study, did not improve over time. The difference in student performance between the two tests is offered by the CRESST researchers as proof that high-stakes tests "narrow the curriculum" and induce "teaching to the test." Test critics would describe the first set of scores as artificially inflated, polluted, or corrupted.

The idea behind score inflation is that as teachers become more familiar with test content, they spend more time teaching that test content and less time teaching other material. So, over time, as familiarity grows, scores climb on the test while real learning suffers.

In the early 1980s, a West Virginia physician named John J. Cannell investigated a statistical anomaly that he had discovered:

Statewide average scores for students on some widely used test batteries were above the national average in every state in which they were given.[73] It was dubbed the Lake Wobegon Effect after the fictional community where "all the children are above average."

Response:

The skeptical reader might see the catch-22. Some of the same critics who argue that tests must be well-aligned to a curriculum in order to be valid will howl "narrowing the curriculum" when scores increase on an aligned test, but not on an unaligned test. There is no justifiable reason why one should expect student test scores on a test not aligned to their curriculum to increase over time. Nor is there anything sacrosanct about the other, unaligned test used in the CRESST researchers' study. One wouldn't expect student scores to increase on a plumbers' test after they had taken an electricians' course, either. But one would be worried if their scores on an electricians' test did not improve.

The Lake Wobegon anomaly might have been caused—observed Cannell and a group of test experts—by a number of factors, including schools' reusing old tests year after year and growing familiar with their specific content and test publishers' waiting years before "renorming" the reference scales. Other factors could have included: the "nonrepresentativeness" of the norming samples;[74] school districts' choosing from among various versions of tests the one most closely aligned with their curriculum and on which their pupils would likely perform best; and the fact that student achievement really was improving throughout the 1980s, as verified by independent testing, such as the scores on SAT, ACT, and NAEP exams. There may also have been some statistical anomalies in Dr. Cannell's calculations.[75]

The Lake Wobegon controversy led to calls for more state control over test content and administration and less local discretion. In most states, those calls were answered. Today most school districts are aware of the problem of test score inflation and do not use tests with exactly the same questions year after year. Many jurisdictions now either use tests that are custom-built to their

state standards and curricula or that are adapted from commercial publishers' test-item banks. A simple way of preventing score inflation is to use different tests or test forms from year to year without announcing in advance which one will be used. Indeed, most of the likeliest sources of the Lake Wobegon effect are fairly easily avoided.[76]

The larger argument about teaching to the test has several components.

2. Curriculum Narrowing

We might suppose that preparing youngsters to do well on tests would find favor with testing experts, yet many of them condemn all forms of "teaching to the test." These arguments tend to come in several forms. One is that valuable subjects that are not tested (for example, art and music, maybe even social studies or science) will be ignored or slighted by test-obsessed teachers and school systems. In her talk at the AERA press conference, for example, Lorrie Shepard of CRESST and the University of Colorado asserted: "Although critics may originally have feared that testing would take instructional time away from 'frills,' such as art and citizenship, the evidence now shows that social studies and science are neglected because of the importance of raising test scores in the basic skills."[77]

A variation on this theme holds that even within a subject that is taught, content coverage will be narrowed (or curricular depth made shallow) in order to conform to the content or style of the test.

Response:

There is only so much instructional time available and choices must be made as to how it is used. (Of course, some new school designs extend the school day or year to ameliorate this problem.) If non-tested subjects are being dropped, either they, too, should be tested or, perhaps, educators and policy makers are signaling that, in a world of tough choices among competing priorities, some subjects must in fact take a backseat to others. A state or school system could easily add high-stakes tests in art,

music, language, and civics, or any other subjects. Attaching high stakes to tests in some subjects and not others would be interpreted by most as a signal that the former subjects are considered to be more important. Perhaps that's actually true. Especially where students are sorely deficient in basic skills and need extra instruction in them, it is likely that few parents would object to such priorities. Survey results show clearly that the public wants students to master the basics skills first, before they go on to explore the rest of the possible curriculum.[78] If that means spending more time on "the basics," so be it. As for subject content being narrowed or made shallow in anticipation of a test, a better response than eliminating the test might be to replace it with one that probes deeper or more broadly.

3a. Emphasis on Lower-Order Thinking in Instruction

In her talk at the AERA press conference, Lorrie Shepard further asserted:

> High-stakes testing misdirects instruction even for the basic skills. Under pressure, classroom instruction is increasingly dominated by tasks that resemble tests. . . . Even in the early grades, students practice finding mistakes rather than do real writing, and they learn to guess by eliminating wrong answers. . . .
>
> In an extensive eighteen-month observational study, for example, Mary Lee Smith and her colleagues found that because of external tests, elementary teachers had given up on [students'] reading real books, writing, and undertaking long-term projects and were filling all available time with word recognition; recognition of errors in spelling, language usage, and punctuation; and arithmetic operations. . . .[79]

Response:

Critics like Smith and Shepard say that intensive instruction in basic skills denies the slow students instruction in the "the neat stuff" in favor of "lower-order thinking."[80] They argue that time for preparing students for high-stakes tests reduces "ordinary instruction." They cannot abide the notion that preparing students for a standardized test could be considered instruction because it is not the kind of instruction that they favor.[81]

Instruction to which teachers may resort to help students improve their scores on standardized tests tends not to be constructivist. It is the type of instruction, however, that teachers feel works best for knowledge and skill acquisition. Teachers in high-stakes testing situations do not deliberately use instructional practices that impede learning; they use those that they find to be most successful.

These testing critics idealize the concept of teachers as individual craftspersons, responding to the unique needs of their unique pupils in unique ways with "creative and innovative" curriculum and instruction.[82] But the most difficult jobs in the world are those that must be created anew every day without any consistent structure and performed in isolation without collaboration or advice. In Public Agenda's research, "teachers routinely complained that teaching is an isolated and isolating experience."[83]

By contrast, teachers in other countries are commonly held to more narrowly prescribed curricula and teaching methods. Furthermore, because their curricula and instructional methods are standardized, they can work together and learn from each other. They seem not to suffer from a loss of "creativity and innovation"; indeed, when adjusted for a country's wealth, teachers in other nations are commonly paid more and usually have greater prestige.[84]

Critics like Shepard and Smith cannot accept that some teachers may *want* to conform to systemwide standards for curriculum, instruction, and testing. Standardization brings the security, convenience, camaraderie, and common professional development that accompany a shared work experience.[85]

3b. Emphasis on Lower-Order Thinking in Test Content

One CSTEEP study, funded by the National Science Foundation, analyzed whether several widely used commercial (and mostly multiple-choice) tests required higher- or lower-order thinking. A press account boasted, "In the most comprehensive study of its kind yet conducted, researchers from Boston College have found evidence to confirm the widespread view that standardized and textbook tests emphasize low-level thinking

and knowledge and that they exert a profound, mostly negative effect on classroom interaction."[86]

Researcher Maryellen Harmon told a reporter, "None of [the test content] calls for high-order thinking that requires that they go in-depth into the concept, that they use math skills in non-conventional contexts, or pull together concepts from geometry and algebra."[87] Project director George Madaus was quoted as saying that the findings present a "depressing picture. . . . If this doesn't change, an inordinate amount of time, attention, and preparation will be given to the wrong domains in math and science, domains that are not reflecting the outcomes we want."[88]

Response:

Many readers would be astonished, as I still am, by the vehemence of some critics' ire toward something as seemingly dull and innocuous as item-response format. And many of the accusations leveled at multiple-choice items have little substance. For example, you can often find in CSTEEP and FairTest publications assertions that multiple-choice items demand only factual recall and lower-order thinking, while performance-based tests do neither. Both claims are without merit. It is the structure of the *question,* not the response format, that determines the character of the cognitive processing necessary to reach a correct answer.

Test items can be banal and simplistic or intricately complex, and either way, their response format can be multiple-choice or open-ended. There is no necessary correlation between the difficulty of a problem and its response format. Even huge, integrative tasks that require fifty minutes to classify, assemble, organize, calculate, and analyze can, in the end, present the test-taker with a multiple-choice response format. Just because the answer to the question is among those provided, it is not necessarily easy or obvious how to get from the question to the answer.

Anyone who still thinks that multiple-choice items demand only factual recall should take a trip to the bookstore and look at some SAT or ACT help books. I purchased a copy of the Cliffs Notes SAT prep book and randomly picked a page. It was in the

math section and four items are posed. Here's one: "What is the maximum number of milk cartons, each 2" wide by 3" long by 4" tall, that can fit into a cardboard box with inside dimensions of 16" wide by 9" long by 8" tall?" Five possible answers are provided, but the correct one, obviously, cannot just be "recalled." Calculations are required. My solution was to calculate the area, in square inches, of a carton and the box, by multiplying the three dimensions in each case, then to divide the former area into the latter. I used pen and paper for two of the calculations and figured the other in my head. Interestingly, the Cliffs Notes book solves the problem graphically, by sketching a three-dimensional box and subdividing it along each dimension.[89]

Indeed, much of the Cliffs Notes book is devoted to convincing the student that there is usually more than one way to "construct" a response to a problem. The book contains sections that illustrate different approaches to solving similar problems. It's a very "constructivist" book; any student following its advice would make ample use—in taking the SAT—of pen, paper, calculator, formulas, diagrams, sketches, lateral thinking, meta-analysis, and other devices that constructivists hold dear. Students armed with multiple methods for solving problems, of course, will hit more correct answers on the SAT than students with fewer methods, other factors held equal. So, higher SAT scores should be taken as evidence of more higher-order thinking.

All the optical scanner will read in the end, however, is a sheet of circles, some filled in with pencil and others not. Moreover, all the computer will score in the end is the number of correct filled-in circles. The calculations, sketches, and diagrams the student used to solve the problems are left behind in the test booklet, on scratch paper, or in the student's head. Just because the optical scanner and computer do not see the process evidence of higher-order thinking, however, does not mean it did not take place.[90] Yet that is what the critics assume.

The most essential point for the critics in applying the "lower-order" label to multiple-choice and the "higher-order" label to performance tests seems to be that with open-ended questions, a student *shows* her work in the test-response book itself and a

scorer can see how the test-taker has approached the problem through the exposition of the answer. This is undoubtedly helpful to teachers but far less necessary for purposes of informing parents, policy makers, admissions counselors, and so on.

CSTEEP's study of several commercially available (and mostly multiple-choice) math and science tests claimed to analyze whether the tests required higher- or lower-order thinking. The researchers defined higher-order thinking skills as having three characteristics: problem solving (the abilities to formulate problems, use a variety of problem-solving strategies in nonroutine situations, and verify and interpret results); reasoning (the abilities to infer, analyze, and use logic); and communicating (the abilities to speak, write, depict, or demonstrate ideas in prose, graphs, models, equations and to describe, explain, or argue a position).[91]

The first two characteristics are typically found in definitions of higher-order thinking.[92] The third was added by CSTEEP for the purposes of their study. The CSTEEP researchers crafted a definition of higher-order thinking that multiple-choice tests would invariably fail. According to the CSTEEP researchers, one is *not* communicating when filling in a bubble for a multiple-choice item, no matter what mental or physical processes may have been used in getting a student to that point; but one *is* communicating when writing a textual response to an open-ended prompt.[93] If the scorer cannot see the work, the work does not exist. Obviously, if one can define higher-order thinking skills any way one wishes, as these CSTEEP researchers did, one can define any type of testing one dislikes as embodying only lower-order thinking.

Even when defined without the communicating component, is higher-order thinking always a superior form of thinking, as testing critics imply? Consider the type of thinking surgeons do. They are highly paid and well-respected professionals. Their course of study, however, consists of a considerable amount of rote memorization, and their work entails a considerable amount of routine and factual recall (all lower-order thinking). Moreover,

the medical college admissions test is largely multiple-choice, and tests administered during medical training largely elicit the recall of discrete facts.

If you were about to go under the knife, which kind of surgeon would you want? One who used only higher-order thinking, only "creative and innovative" techniques, and "constructed her own meaning" from every operation she performed?

Or would you prefer a surgeon who had passed her lower-order thinking exams—on the difference, say, between a spleen and a kidney—and used tried-and-true methods with a history of success, methods that other surgeons had used successfully?

Certainly, there would be some situations where one could benefit from an innovative surgeon. If *no aspect whatsoever* of the study or practice of surgery were standardized, however, there would be nothing to teach in medical school and your regular barber or beautician would be as well qualified to "creatively" excise your appendix as anyone else. Ideally, most of us would want a surgeon who possesses both lower and higher abilities.[94]

The surgery analogy also addresses another of the testing critics' arguments. They say that multiple-choice tests limit students to the "one correct answer" when there may be more than one valid answer and more than one way to get to each. Moreover, they say, students should not get an entire exercise counted wrong if they analyze most of the problem correctly, but make one careless error.

Most of us would sympathize with this sentiment, but we should remember that there are countless examples in real life where there *is* just one right answer or where one careless error can have devastating consequences—in brain surgery, for example.

4. Declining Achievement

Testing experts claim that high-stakes tests actually interfere with learning and student achievement in states that use them. In "High-Stakes Tests Do Not Improve Student Learning," FairTest asserted that states with high-stakes graduation exams

tend to score lower on the NAEP. According to FairTest, this "contradicts the . . . common assumption of standards and tests-based school reform . . . that high-stakes testing . . . will produce improved learning outcomes."[95]

The FairTest solution is to restrict testing to occasional no-stakes monitoring with samples of students using the types of response formats that FairTest favors (no multiple-choice). Scores on "portfolios" of each student's best work would track individual student progress.[96] Indeed, the only state-testing program to garner the highest rating from FairTest was Vermont, which had a statewide portfolio program and no high-stakes or multiple-choice standardized testing.[97]

Response:

The claim that high-stakes tests inhibit learning is a weak argument supported by dubious research. The FairTest report provides a good example of just how simplistic that research can be. FairTest argues that states with high-stakes, minimum-competency-test graduation requirements tend to have lower average test scores on the NAEP. They make no effort, however, to control for other factors that influence test performance, and the relationship between cause and effect is just assumed to run in the direction FairTest wants.[98] Most honest observers would assume the direction of cause and effect to be just the opposite—poorly performing states initiate high-stakes testing programs in an effort to improve academic performance while high-performing states do not feel the need to.

The work of Cornell labor economist John Bishop does not get the press attention bestowed on FairTest. Yet in a series of solid studies conducted over a decade, Bishop has shown that when other factors that influence academic achievement are controlled for, students from states, provinces, or countries with medium- or high-stakes testing programs score better on neutral, common tests and earn higher salaries after graduation than do their counterparts from states, provinces, or countries with no- or low-stakes tests.[99]

Bishop recently turned his attention to the very same relationship that FairTest studied, except he looked at it in depth. He and his colleagues used individual-level data from the National Education Longitudinal Study that began in 1988 (NELS:88) and High School and Beyond (HSB), another longitudinal study that ran from 1980 to 1992. They controlled for socioeconomic status, grades, and other important factors, while comparing the earnings of graduates from "minimum-competency" testing states to those from non-testing states.[100] "They found that test-taking students earned an average of 3 percent to 5 percent more per hour than their counterparts from schools with no minimum-competency tests. And the differences were greater for women, with as much as 6 percent higher earnings for those who had taken the tests. Other evidence of the success of high-stakes state testing programs continues to surface.[101]

5. Standardized Tests Hurt Women and Minorities

As mentioned in the case study of high-stakes testing in Texas, the NAACP and the Mexican-American Legal Defense Fund both argued that the Texas Assessment of Academic Skills was biased against minorities.

The brunt of FairTest's attack on the SAT involves alleged bias as well. The argument is straightforward: On average, girls score worse on the SAT than boys, despite getting better grades in school. Therefore the SAT is gender-biased. Blacks and Hispanics score lower than whites. Therefore the SAT is race-biased.[102] FairTest argues that these biases depress minority and female college admissions.

Response:

After investigating why girls score worse on the SAT than boys despite getting better grades in school, the Educational Testing Service (ETS), the SAT's developer, concluded that the gender difference in SAT scores was almost entirely explained by high school course selection (for example, girls took fewer math and science courses than boys and so got lower SAT math scores).[103]

FairTest called the ETS explanation a "smokescreen."[104] Yet similar evidence is available for blacks and Hispanics: Almost all the SAT math score differences between them and their white counterparts disappear when they take as much algebra and geometry in high school as white students do.[105]

The charge that the use of SATs in college admissions artificially depresses minority admissions is also misguided. As David W. Murray writes in "The War on Testing":

> Nor is it even clear that relying more exclusively on grades would bump up the enrollment numbers of blacks and Hispanics, as many seem to think. While it is true that more minority students would thereby become eligible for admission, so would other students whose grade point averages (GPAs) outstripped their test scores. A state commission in California, considering the adoption of such a scheme, discovered that in order to pick students from this larger pool for the limited number of places in the state university system, the schools would have to raise their GPA cut-off point. As a result, the percentage of eligible Hispanics would have remained the same, and black eligibility actually would have dropped.[106]

There is a double sadness to the focus of some minority spokesmen on the messenger instead of the message. Black and Hispanic students in the United States generally receive an education inferior to what white students receive. This is a shame and a disgrace. By blaming standardized tests instead of the schools that are responsible for their students' poor achievement, however, these advocacy groups waste efforts that would be better expended reforming bad schools.

A Public Agenda survey of parents on education issues pertaining to race implies that the NAACP actions in Texas and other states against high-stakes standardized testing may not even reflect what most African Americans want. "Most African-American parents do not think standardized tests are culturally biased," reports Public Agenda, "and very few want race to be a factor when choosing the best teachers for their children . . . ."[107] When asked why, on average, black students don't do as well as whites on standardized achievement tests, only 28 percent say it is mostly because "the tests are culturally biased against black stu-

dents." Forty-four percent of black parents say "the tests measure real differences in educational achievement," and 18 percent say the reason for this difference is a result of low expectations.[108]

6. Tests Are Too Costly

Some experts have criticized standardized tests as too costly. Daniel Koretz appeared before a Congressional committee to testify against President George H. W. Bush's national testing proposal and stressed cost as a major negative. He claimed that while costs of standard multiple-choice commercial tests range from $2 to $5 per student, the costs of performance-based national tests would be considerably more than $100 per student, perhaps as high as $325 per student.

Another study of the extent and cost of testing, by Walter Haney and George Madaus of CSTEEP, calculated a high estimate of $22.7 billion spent on standardized testing in a year.[109] U.S. schools, the CSTEEP report claimed, suffered from "too much standardized testing" that amounted to "a complete and utter waste of resources."[110] Their estimate breaks down to about $575 per student per year.

A CRESST report by Larry Picus, which counted cost components in much the same way as the CSTEEP study, estimated costs of a certain state test at between $848 and $1,792 per student tested.[111]

Response:

In the early 1990s, the U.S. General Accounting Office (GAO) surveyed a national sample of state and local testing directors and administrators to appraise the costs of then-current statewide and districtwide tests. Based on their responses, the GAO assumed that the tests President George H. W. Bush proposed would probably evolve, as many state exams had, from plans for a 100 percent performance-based format to a mixed format that includes multiple-choice items. Eleven state tests ranging from 20 percent to 100 percent performance-based cost an average of $33 per student, including the salary time of teachers and other staff engaged in test-related activity, as well as the purchase of test

materials and services. The GAO estimated that slightly more than $500 million was spent by U.S. school systems on systemwide testing in a year, or about 0.2 percent of all spending on elementary and secondary schools.[112]

The GAO estimate of $33 per student contrasts with CRESST and CSTEEP estimates of $575 to $1,792. The GAO estimate of about $500 million for the total national cost of systemwide testing contrasts with a CSTEEP estimate 45 times higher.

Testing critics estimate standardized tests' costs so much higher because they count the costs of any activities *related to* a test as costs *of* a test. In the CRESST study of Kentucky's performance-based testing program, for example, teachers were asked to count the number of hours they spent preparing materials for classroom use that related to the testing program. In an instructional program that has the intention of unifying all instruction and assessment into a seamless web, where the curriculum and the test mutually determine each other, all instruction throughout the entire school year will be related to the assessment.

Furthermore, the Kentucky Instructional Results Information System (KIRIS) was a comprehensive program that included changes in curriculum, instruction, and evaluation. Assessment was just one component. All the changes were implemented at the same time, and some survey respondents could consider any or all KIRIS costs as related to the assessment. Given the manner in which it posed its questions, CRESST could not discern which were costs of the test and which were costs of other parts of the KIRIS program.

The CSTEEP study counted even more cost items, such as student time. Walter Haney and the other CSTEEP researchers assumed that the time spent on preparing for or taking a test holds no instructional value whatsoever. (I would guess that students probably learn more while preparing for or taking a test.) Then they calculated the present discounted value of that "lost" instructional time against future earnings, assuming all future earnings to be the direct outcome of school instruction. The

CSTEEP researchers also counted building overhead (maintenance and capital costs) for the time spent testing, even though those costs are constant and not affected by the existence of a test. In sum, CSTEEP counts any and all costs incurred simultaneously with tests, not just those caused by testing.

7. Other Countries Don't Test As Much As the United States

At the AERA press conference, CSTEEP's Madaus argued against proposals for a national examination system by claiming that "American students [were] already the most heavily tested in the world."[113] In a separate report, he also asserted that the trend in other developed countries is toward less standardized testing. He reasoned that other countries are dropping large-scale external tests because they no longer need them as selection devices because places in upper secondary programs are being made available to everyone and access to higher education programs has widened. Thus, he argued, a worldwide trend toward less external testing could be found at all levels of education, "even at the postsecondary level," and it was unidirectional—large-scale, external tests were being "abolished."[114]

Response:

Are U.S. students the "most heavily tested in the world"? No. U.S. students actually spend less time taking high-stakes standardized tests than do students in other developed countries. A 1991 survey for the Organisation for Economic Co-operation and Development revealed that "U.S. students face fewer hours and fewer numbers of high-stakes standardized tests than their counterparts in every one of the thirteen other countries and states participating in the survey and fewer hours of state-mandated tests than their counterparts in twelve of the thirteen other countries and states."[115]

What of Madaus's assertion of a trend toward less standardized testing in other countries?[116] The primary trend appears to be toward more testing, with a variety of new test types used for a variety of purposes. In a study I conducted, I found twenty-seven

countries and provinces had increased or planned to increase testing over the period 1974–1999, while only three decreased it. Altogether, fifty-nine tests were added and only five dropped.[117]

8. All Those Who Really Care about Children Oppose Testing

The panelists at the AERA press conference implied that they were speaking on behalf of teachers and students, defending them against politicians, mean-spirited conservatives, and the greedy testing industry.[118] The critics claimed that those who care about teachers and students see testing for what it really is and oppose it.

Regarding teachers, for example, Robert Stake, said, "[teachers] have essentially no confidence in testing as the basis of the reform of schooling in America."[119]

The laundry list of costs attributed to *students* from the use of standardized tests ranged from a change in instruction away from the "neat stuff" in the curriculum toward "lower-order thinking" to an increase in grade retention and dropout rates from the use of standardized tests in high-stakes situations. A CRESST study on the "unintended consequences of external testing" that Mary Lee Smith referred to at the conference claimed to find "stress, frustration, burnout, fatigue, physical illness, misbehavior and fighting, and psychological distress" among the effects of testing on young students.[120]

Response:

To learn the true attitudes toward testing among teachers, students, parents, and the public, I attempted to gather all relevant U.S. poll and survey items on student testing by collecting many surveys myself and searching the Roper Center archives. I discovered 200 items from seventy-five surveys over three decades.[121]

The results are fairly decisive. Majorities of the general public favor more testing, more high-stakes testing, and higher stakes in testing. The majorities have been large, often very large, and fairly consistent over the years, across polls and surveys, and even across respondent groups. Parents, students, employers, state education

administrators, and even teachers (who exhibit more guarded opinions and sometimes fear being blamed if their students score badly on tests) consistently favor more student testing and higher stakes.

Twenty-seven polls taken between 1970 and 1999 asked specific respondents whether they thought education would improve if there were higher (student) stakes in school testing. The results of twenty-six of the twenty-seven polls said yes, in most cases by huge margins.

Which was the twenty-seventh study, the one claiming that respondents want lower stakes in student testing? It was a survey conducted by CSTEEP and funded by the National Science Foundation.[122] Its contrary conclusions may have a lot to do with its convoluted design. First, respondents were chosen selectively from urban, high-minority public school districts. High school teachers in the sample were limited to those with classes of "average and below average" students.[123] Moreover, the specific interview question that elicited opinions on the effects of mandated tests was, in my judgment, biased in a way that would generate negative answers. The question was: "Do you have any particular concerns or opinions about any of these standardized tests?" "Concerns" doesn't equal "criticisms" in meaning, but, in this context, it's pretty close.[124] Then the CSTEEP researchers classified as "negative" responses those that others might classify as neutral or positive. For example, if a teacher said that her students "didn't test well," it was interpreted by the researchers as a "major source of invalidity" and a "negative" comment, even though students can test poorly for dozens of reasons, including not studying and not paying attention in class.[125]

Do these CSTEEP researchers and the speakers at the aforementioned AERA press conference at least represent other "education establishment" organizations in opposing high-stakes standardized testing?

Far from it. The National Association of State Boards of Education has come out strongly in favor of greater use of high-stakes standardized testing.[126] So have state superintendents and governors. The American Federation of Teachers (AFT) has

been the nation's most forceful and vocal advocate for greater use of high-stakes standardized student testing.

Nationwide polls of teachers conducted over three decades by the Carnegie Foundation for the Advancement of Teaching, Metropolitan Life Insurance Co., the AFT, and Public Agenda show strong teacher support for high-stakes standardized tests.[127]

Despite this widespread support for testing, press coverage of testing issues often seems one-sided *against* testing. It typically features a FairTest spokesperson as the antitesting alternative to some sincere, beleaguered state or local testing director just trying to do her job.

I telephoned a few newspaper reporters to try to understand why their stories on testing were set up this way. They replied that they do not know of any advocacy group on the other side of the issue that could balance FairTest's perspective. They added that FairTest is also reliable: They keep up with the issues, and they return telephone calls promptly. In his review of SAT critiques, Gregory Cizek expresses disappointment that "the measurement profession has made no corresponding, popular, accessible, public defense of its mission or of testing."[128]

While one sees only a handful of education-researcher experts speaking out in favor of high-stakes standardized tests, there are in fact hundreds of qualified testing experts working for national, state, or local agencies (not to mention the experts working for organizations that develop tests under contract to these government agencies) who are legally and ethically restricted from expressing their views regarding testing policy. The debate seems unbalanced only because one side is often missing from it, that of the pro-testing advocates who cannot speak out.

For a reporter who arrives at the office in the morning with no story and who cannot leave in the evening without one, FairTest is a godsend. The millions who favor testing have no comparable voice.

Testing in Perspective

That tests and test results can be misused is beyond dispute. Human beings are responsible for administering them and inter-

preting their results, and humans are imperfect creatures. There also is no denying that tests are imperfect measurement devices. If the items in the antitesting canon were also beyond dispute, one might well be disposed to give up on high-stakes standardized testing. But that would be an enormous mistake.

The critics would have us believe that all problems with high-stakes and standardized testing must always be with us, that is, that nothing can be changed or improved. They're wrong. Some of the alleged problems—that they hurt learning and are expensive, for example—are really not problems at all, as shown above. Other problems apply equally to the alternatives to testing. Still others are solvable and are being or have been solved by state, local, or national testing directors.

Probably the single most important recent innovation in relation to the quality and fairness of testing in the United States has been the addition of managerial and technical expertise in state education agencies. At that level, it is possible to retain an adequate group of technically proficient testing experts, adept at screening, evaluating, administering, and interpreting tests, who are not controlled by commercial publishers or naive about test results. They, along with governors and legislatures, are currently calling the shots in standardized testing. Some of the most important decisions affecting the design and content of the tests, the character of the testing industry, and the nature of its work are today being made by state testing directors.

These testing directors can, for example, deploy a number of relatively simple solutions to the problems of score inflation, curricular compression and teaching to the test, including: not revealing the contents of tests beforehand; not using the same test twice; including items on the test that sample broadly from the whole domain of the curriculum tested; requiring that nontested subjects also get taught (or testing them, too); and maintaining strict precautions against cheating during test administrations.

In North Carolina and other states, they do something else about score inflation: They keep raising the bar. As instruction and learning improve and scores rise, they boost their grading standards.[129] Their students' dramatic improvements on the independent NAEP offer evidence that the achievement gains

are real, not a result of score inflation caused by narrowing the curriculum and teaching to the test.

In some states and countries, officials use "blended" or "moderated" scores for high-stakes decisions. The "blends" combine test scores with other measures, such as classroom grades and attendance records, so that instructional efforts will not focus exclusively on the standardized test and so that high-stakes decisions will not be based solely on single or even multiple attempts at passing a test.

One final argument against testing, the argument that using test results to evaluate schools leads to unfair comparisons between rich districts with highly educated parents and poor districts with less-well educated parents, can also be dealt with. There are at least two solutions to this problem. One is to set targets for schools based on their own past performance. Another is to calculate value-added test scores, as Tennessee and North Carolina do. This method estimates how much value a school adds to the level of achievement that would have been predicted (given the background and prior attainment of students), then adjusts a school's or district's test scores accordingly. Like any other system, value-added scoring can be abused; there's a particular danger in its being used to excuse the performance of school systems that have a large number of poor and minority children. Value-added scores can also be tricky to calculate. But many able and earnest analysts throughout the country are striving to make value-added systems work.

Although some of the "problems" with standardized testing turn out not to be problems, and others turn out to be solvable, a third set of problems is inherent and inevitable—but similar problems are also present in the alternatives to standardized tests.

The critics unfairly compare high-stakes standardized testing with their own notion of perfection. Administration of high-stakes tests will never be perfect. There will always be some teachers and pupils who cheat. There will always be some students who are better prepared to take a test than others, and so on.

Perfection, however, is not a reasonable standard of comparison for standardized testing. Too often, the alternative is a system

of social promotion with many levels of (nominally) the same subject matter being taught, ranging from classes for the self-motivated kids to those for youngsters who quit trying years before and whom the system has ignored ever since.[130] Too often, the result is a system that graduates functional illiterates.

If *none* of the curriculum is tested, we cannot know if any of it works. Without standardized tests, no one outside the classroom can reliably gauge student progress. No district or state superintendent. No governor. No taxpayer. No parent. No student. Each has to accept whatever the teacher says, and without standardized tests, no teacher has any point of comparison, either.

Certainly, it is unfair to test what has not been taught, but no such claim can be made about testing what has been taught. And if what is tested is the curriculum, then attacks on teaching to the test seem silly because teachers are teaching what they should be teaching.

Eliminating high-stakes standardized testing would necessarily increase our reliance on teacher grading and testing. But are teacher evaluations free from all the complaints of the antitesting canon? Not exactly. Individual teachers also can narrow the curriculum to that which they prefer. Grades are susceptible to inflation with any teachers. Students get to know a teacher better and learn his idiosyncrasies. A teacher's grades and test scores are far more likely to be idiosyncratic and far less likely to be generalizable than the scores of any standardized test.[131]

Moreover, teacher-made tests are not necessarily any better-supplied with higher-order thinking than are standardized tests. Yet many test critics would bar all high-stakes standardized tests and have us rely solely on teacher evaluations of student performance. How reliable are those evaluations? Not very. There are a number of problems with teacher evaluations, according to research on the topic. Teachers tend to consider "nearly everything" when assigning marks, including student class participation, perceived effort, progress over the period of the course, and comportment, according to Gregory Cizek. Actual achievement vis-à-vis the subject matter is just one factor. Indeed, many teachers express a clear preference for noncognitive outcomes such as "group interaction, effort, and participation" as more important

than averaging tests and quiz scores.[132] It's not so much what you know, it's how you act in class. Being enthusiastic and group-oriented not only gets you into the audience for television game shows, but it also, apparently, gets you better grades in school.

One study of teacher grading practices discovered that 66 percent of teachers feel that their perception of a student's ability should be taken into consideration in awarding the final grade.[133] Parents of students who assume that their children's grades represent subject matter mastery might well be surprised.

Conclusion: Two Views of Testing and Learning

There is perhaps no more concise exposition of the general philosophy undergirding opposition to standardized testing among education experts than that revealed in the Public Agenda survey of education school professors, *Different Drummers*.[134] Among the reasons most dislike standardized tests are their preferences for: "process over content"; "facilitating learning" rather than teaching; and "partnership and collaboration" over imparting knowledge.[135]

A large majority of education school professors surveyed felt that it was more important that "kids struggle with the process of trying to find the right answers" (86 percent) than that "kids end up knowing the right answers to the questions or problems" (12 percent): "[I]t is the process, not the content, of learning that most engages the passion and energy of teacher educators. If students learn how to learn, the content will naturally follow."[136]

The role of teachers in this education worldview, then, should be that of "facilitator," not "sage on the stage." When asked which statement was "closer to their own philosophy of the role of teachers," 92 percent of the education professors agreed that "teachers should see themselves as facilitators of learning who enable their students to learn on their own." Only 7 percent felt that "teachers should see themselves as conveyors of knowledge who enlighten their students with what they know."[137]

The constructivist criticism of any teaching or testing that fixes the manner of solving a problem and penalizes students for care-

less or minor errors is not shared by the public or even by students. In *Getting By,* Public Agenda reported that 79 percent of teens say "most students would learn more if their schools routinely assured that kids were on time and completed their homework . . . . [Sixty-one percent said] having their classwork checked regularly and being forced to redo it until it was correct would get them to learn a lot more. When interviewed in focus groups, teens often remembered "tough" teachers with fondness: "I had a math teacher [who] was like a drill sergeant. She was nice, but she was really strict. Now I don't have her this year, and looking back, I learned so much."[138]

In the real world, testing will continue. Testing experts have much to contribute to efforts that ensure testing is done well. Unfortunately, many of them share an ideological orientation that makes any type of standardized test impossible to swallow. Until these experts reexamine their most fundamental beliefs about teaching and learning, all the hard work of improving standardized tests will have to be done without them.

Notes

1. See Richard P. Phelps, "The Demand for Standardized Student Testing," *Educational Measurement: Issues and Practice* 17, no. 3 (fall 1998).
2. Steve Farkas, Jean Johnson, and Ann Duffett, *Different Drummers: How Teachers of Teachers View Public Education* (New York: Public Agenda, 1997), 20, 36.
3. Ibid., 20.
4. Ibid., 13–14.
5. Ibid., 13.
6. Ibid., 14.
7. See transcripts of the conference papers printed, along with an introduction, in "Accountability As a Reform Strategy," *Phi Delta Kappan* (November 1991): 219–51.
8. CRESST is headquartered at UCLA's education school, but associates with "partners" at the education schools of the universities of Colorado and Southern California, Arizona State University, and the RAND Corporation. CRESST publishes dozens of reports every year that are objective, often concentrating on psychometric methods. Some of the best psychometric research in the country is

produced by CRESST. The research that several of CRESST's affiliated scholars publish that relates to *testing policy,* however, typically subscribes to the canon.

9. CSTTEEP has recently changed its name to the National Board on Educational Testing and Public Policy (NBETPP). Not everyone associated with CSTEEP opposes testing. Indeed, the Third International Mathematics and Science Study (TIMSS), perhaps the standardized test most reviled by testing critics, was headquartered at CSTEEP. TIMSS showed U.S. students performing more and more poorly in comparison with their international counterparts as grade levels advanced to the last year of high school. But another group of researchers at CSTEEP, not associated with TIMSS, devote themselves almost exclusively to antitesting research.
10. FairTest receives much of its financial support from the Ford Foundation, a great deal of exposure in the media, and a seat at the table with study commissions on testing policy as an interested "stakeholder"; for example, it was on the former Office of Technology Assessment's Advisory Panel for the report *Testing in American Schools: Asking the Right Questions.*
11. The actual number of standardized tests administered annually in the public schools at the time was around 40 million, not 100 million. Forty million tests for about 40 million students calculates to one test per student per year, and only one in four of those was for high stakes. See Richard P. Phelps, "The Extent and Character of Systemwide Student Testing in the United States," *Educational Assessment* 4, no. 2: 89–121.
12. Noe Medina and Monty D. Neill, *Fallout from the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools* (Cambridge, Mass.: FairTest, 1990).
13. Daniel M. Koretz, "State Comparisons Using NAEP: Large Costs, Disappointing Benefits," *Educational Researcher* 20, no. 3 (April 1991): 19.
14. Ibid., 19–21.
15. Ibid., 20.
16. See also Richard M. Wolf, "What Can We Learn from State NAEP?" *Educational Measurement: Issues and Practice;* Bruce J. Biddle, "Foolishness, Dangerous Nonsense, and Real Correlates of State Differences in Achievement," *Phi Delta Kappan* (Bloomington, Ind.: Phi Delta Kappa online article, August 8, 1998).
17. Koretz, "State Comparisons Using NAEP: Large Costs, Disappointing Benefits," 20.
18. Gary Phillips, "Benefits of State-by-State Comparisons," *Educational Researcher*

20, no. 3 (April 1991), 17–19.

19. Ibid., p. 17. See also George Bohrnstedt, Project Director, *The Trial State Assessment: Prospects and Realities* (Stanford, Calif.: National Academy of Education, 1993).
20. James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress* (Washington, D.C.: National Academy Press, 1998), 2.
21. Ibid.
22. FairTest, "How the States Scored" (Cambridge, Mass.: FairTest, summer 1997).
23. Robert C. Johnston, "In Texas, the Arrival of Spring Means the Focus Is on Testing," *Education Week* 17, no. 33 (April 29, 1998): 20.
24. Ibid., 21.
25. Lonnie Harp, "OCR Probes Bias Complaint Against Texas Exit Test," *Education Week on the Web* (February 7, 1996).
26. Linda Jacobson, "State Graduation Tests Raise Questions, Stakes," *Education Week on the Web* (June 24, 1998).
27. "ED Clears Texas Tests" in "News in Brief," *Education Week on the Web* (August 6, 1997).
28. Johnston, "In Texas, the Arrival of Spring Means the Focus Is on Testing," 21.
29. Ibid., 1, 20, 21; "Pass or Fail," *Teacher Magazine: Education Week on the Web* (September 1994); Lonnie Harp, "Final Exam," *Teacher Magazine: Education Week on the Web* (September 1994); Lonnie Harp, "Texas Politicians Wrangle Over School Rankings," *Education Week on the Web* (September 14, 1994); Robert C. Johnston, "Texas Governor Has Social Promotions in His Sights," *Education Week on the Web* (February 11, 1998).
30. See "Pass or Fail," *Teacher Magazine: Education Week on the Web* (September 1994); Harp, "Final Exam"; Harp, "Texas Politicians Wrangle Over School Rankings"; Johnston, "Texas Governor Has Social Promotions in His Sights."
31. They have eased one requirement, however. The "no pass, no play" rule, originally recommended by an education reform commission chaired by Ross Perot, barred students who were failing courses from participating in team sports for six weeks. That has been reduced to three weeks as other, broader requirements have been put in place. See Harp, "Texas Politicians Wrangle Over School Rankings"; Lonnie Harp, "Texas Lawmakers Reach Accord on Overhaul of Education Laws," *Education Week on the Web* (May 31, 1995).
32. See Kathleen Kennedy Manzo: "N.C. Consensus Pushes for New Set of Reforms," *Education Week on the Web* (April 9, 1997); "Quality Counts, '98: North

Carolina Summary," *Education Week on the Web;* "High Stakes: Test Truth or Consequences," *Education Week,* 17, no. 8 (October 22, 1997); "N.C. Gets First School-by-School Performance Results," *Education Week* (September 3, 1997): 26; "Struggling N.C. Schools Buoyed by State Teams," *Education Daily* (July 10, 1998): 4.

33. Marzo, "Struggling N.C. Schools Buoyed by State Teams," 4.
34. Ibid., 4.
35. Telephone conversations with Vanessa Jeter and Janet Byrd of the North Carolina Department of Public Instruction (October 19, 1998).
36. David Molpus, "Improving High School Education," *National Public Radio Morning Edition* (September 15, 1998).
37. Ibid.
38. Manzo, "High Stakes: Test Truths or Consequences," 1, 2.
39. Ibid., 3.
40. Ibid., 3.
41. Ibid., 2.
42. Ibid., 1.
43. Ibid., 2.
44. Ibid., 4.
45. Telephone conversations with James Causby, superintendent of the Johnson County Schools (September 24, 1998).
46. Richard M. Jaeger, "Legislative Perspectives on Statewide Testing" in "Accountability As a Reform Strategy," *Phi Delta Kappan* (November 1991): 242.
47. Ibid.
48. Ibid.
49. George F. Madaus, "The Effects of Important Tests on Students: Implications for a National Examination System," *Phi Delta Kappan* (November 1991): 228; Lorrie A. Shepard, "Will National Tests Improve Student Learning?" *Phi Delta Kappan* (November 1991): 234.
50. See Indicator C3 in Organisation for Economic and Co-operative Development, "Education at a Glance: OECD Indicators 1997" (Paris: OECD, 1997). The United States has a lower percentage of 16-year-old students enrolled than do Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, Germany, the Netherlands, Norway, New Zealand, and Sweden, all countries with high-stakes secondary-level exit exams. Rates in Hungary and Ireland are similar to ours. Switzerland and the United Kingdom are the only countries, among those

included, with high-stakes exit exams and lower enrollment rates than the United States. Comparisons at age 17 are similar. The conclusion: Students drop out in the United States for reasons other than not passing an exit exam.

51. J. B. Erickson, *Indiana Youth Poll: Youths' Views of High School Life* (Indianapolis: Indiana Youth Institute, 1991), 33.
52. Bryan W. Griffin and Mark H. Heidorn, "An Examination of the Relationship Between Minimum Competency Test Performance and Dropping Out of High School," *Educational Evaluation and Policy Analysis* 18, no. 3 (fall 1996): 243–52.
53. C. Boyden Gray and Evan J. Kemp Jr., "Flunking Testing: Is Too Much Fairness Unfair to School Kids?" *Washington Post,* 19 September 1993, C3.
54. FairTest, "Testing Our Children: North Carolina" (Cambridge, Mass.: FairTest, summer 1997).
55. See Indicators 8 and 9 in U.S. Department of Education, National Center for Education Statistics, *State Indicators in Education 1997,* NCES 97–376, by Richard P. Phelps, Andrew Cullen, Jack C. Easton, and Clayton M. Best, Project Officer, Claire Geddes (Washington, D.C.: 1997).
56. While the SAT tends to be more visible, about half of state colleges and well over one-third of all U.S. colleges use the competing American College Test (ACT). Some colleges allow applicants to submit either test.
57. FairTest, "FairTest Fact Sheet: The SAT" (Cambridge, Mass.: FairTest, Summer 1997): 2.
58. They don't mention, however, the continual *growth* in the number of colleges *using* the SAT—more than a hundred added since 1990, bringing the total to 1,450 four-year institutions. See Charles A. Kiesler, "On SAT Cause and Effect," *Education Week* (May 13, 1998): 43.
59. See Debbie Goldberg, "Putting the SAT to the Test," *The Washington Post Education Review* (October 27, 1996): 20–21; Many colleges have complained to the National Association for College Admission Counseling (NACAC) about their presence on "The List" of colleges that FairTest claims waive the SAT requirement. Many colleges FairTest includes waive the requirement only for a few students under extraordinary circumstances (for example, disabilities, remote foreign locations, and so on) (telephone conversation with NACAC officials, August 14, 1998).
60. See, for example, Joyce Slayton Mitchell, "A Word to High School Seniors—SATs Don't Get You In," *Education Week* (May 29, 1998): 33; also National Association for College Admission Counseling, "Members Assess 1996

Recruitment Cycle in Eighth Annual NACAC Admission Trends Survey," *News from National Association for College Admission Counseling* (October 28, 1996): 2, 4.

61. National Association for College Admission Counseling, "Members Assess 1996 Recruitment Cycle in Eighth Annual NACAC Admission Trends Survey," 2, 4.
62. See Warren W. Willingham et al., *Predicting College Grades* (New York: The College Board, 1990), chapters 5 and 12; also Thomas F. Donlon, ed., *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests* (New York: The College Board, 1984), chapter 8.
63. Actually, one could make the same (poor) argument about high school grade-point averages. After all other predictive factors are accounted for, including SAT or ACT scores, high school GPA explains only another several percentage points of the variation in first-year college grades. So, why bother with the GPA? See also Gerald W. Bracey, "The $150 Million Redundancy," *Phi Delta Kappan* 70, no. 9 (May 1989): 698–702; Lucy May, "Tests Don't Have All the Answers to How Kentucky Kids Rank," *The Lexington Herald-Leader,* July 6, 1995; Edward B. Fiske, "Questioning an American Rite of Passage: How Valuable Is the SAT?" *New York Times*, January 18, 1989, B10. Walt Haney of Boston College is often quoted on this issue. See, for example, May, "Test Don't Have All the Answers to How Kentucky Kids Rank"; Debbie Goldberg, "Putting the SAT to the Test"; Peter Sacks, "Standardized Testing: Meritocracy's Crooked Yardstick," *Change* (March/April 1997): 26.
64. Lucy May, "Tests Don't Have All the Answers to How Kentucky Kids Rank," 44, 45.
65. Even 15 percent underestimates the predictive power of the SAT or ACT. Because colleges publish the mean and range of the admissions test scores of their first-year class, a high school senior can pick potential colleges in whose range his test score fits, and he is more likely to be admitted. Likewise, colleges can focus their recruiting efforts where they are likely to find attractive applicants who can succeed on their campuses and who might be willing to come. This represents an added benefit of the SAT—applicants and colleges don't waste time chasing after poor matches. Technicians call this benefit "restriction of range" or, more generally, "allocative efficiency." Allocative efficiency is very difficult to estimate, but the Educational Testing Service has calculated that just for the colleges alone it must add at least another two percentage points of predictive power to the additional 15 percent already accounted for by SAT scores.
66. At best, a student's high school record explains only 44 percent of the variation

in first-year college grades. SAT scores alone explain 42 percent of the variation. To a large extent, however, high school record and SAT scores represent the same thing, mastery of academic subject matter. Thus when high school record and SAT scores are used together in equations to predict students' first-year college grades, the two predictive factors overlap. If SAT scores are put in the equation first, high school record adds only a comparatively smaller amount of predictive power. After subtracting the proportion of predictive power that the two predictive factors share in common, SAT scores predict an additional 15 percent of the variation in first-year college grades. This 15 percent predicted by the SAT represents 34 percent of the variation in first-year college grades explained by high school record alone (which was 44 percent). Thus, if we put high school record in the prediction equation first, SAT scores represent a 34 percent incremental increase in predictive power when added to the equation. See Willingham et al., *Predicting College Grades,* chapters 5 and 12.

67. The annual survey by the National Association for College Admission Counseling shows that their members consider the following criteria the most important in determining admission (by percentage, mentioning each criterion of considerable or moderate importance): grades in college prep courses, such as Advanced Placement courses (90); admission test scores, such as the SAT or ACT (82); grades in all subjects (79); class rank (71); essay or writing sample (53); counselor recommendation (66); teacher recommendation (55). (National Association for College Admission Counseling, "Members Assess 1996 Recruitment Cycle in Eighth Annual NACAC Admission Trends Survey," 2, 4.)
68. To some extent, the criticisms are tautological. A CSTEEP study of several commercially available math and science tests, managed by George Madaus and funded by the National Science Foundation, concluded that the tests promoted "test preparation" practices. Eighty-one percent of math teachers and 53 percent of science teachers engaged in some form of "test preparation," according to CSTEEP. However, the researchers "coded 'test preparation' as 'present' when the teacher or administrator made an explicit link between a particular activity and test scores, or gave such evidence in spite of denying test preparation." Thus, if a teacher taught an ordinary math or science lesson and hoped that it would improve students' performance on a test, that's "test preparation." Mary Maxwell West and Katherine A. Viator, *The Influence of Testing on Teaching Math and Science in Grades 4–12: Appendix D: Testing and Teaching in Six Urban Sites* (Boston: CSTEEP, October 1992), 27–28.
69. Mary Lee Smith et al. *Reforming Schools by Reforming Assessment: Consequences of*

the Arizona Student Assessment Program (ASAP): Equity and Teacher Capacity Building, CSE Technical Report 425 (Los Angeles: CRESST, March 1997), 2.

70. The history of the large-scale, standardized use of portfolios is spare and brief (FairTest, "Testing Our Children: Introduction" [Cambridge, Mass.: FairTest, 1998], 2). There appear to be many problems with a sole reliance on portfolios to measure student progress: They're far more susceptible to cheating, coaching, gaming, and outright plagiarism than are standardized tests. (See "Test Violations Uncovered" in "News in Brief," *Education Week on the Web* [August 6, 1997]: 5; Maryl Gearhart and Joan L. Herman, "Portfolio Assessment: Whose Work Is It?" *Evaluation Comment* [CSE, CRESST, winter 1996]; and Daniel M. Koretz, "Sometimes a Cigar Is Only a Cigar" in *Debating the Future of American Education,* ed. Diane Ravitch [Washington, D.C.: Brookings Institution, 1995]: 160–62.) Moreover, they reward occasional, exceptional brilliance and not steady competence, and they are difficult to score with consistency (Daniel Koretz et al., *The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program,* Technical Report No. 355 [Los Angeles: CRESST, December 1992). These sound like the same tenor of criticisms that FairTest, the most prominent advocate for the exclusive use of portfolios, makes of standardized tests.
71. One can also find elements of other theories and philosophies in the critics' rhetoric—that of "multiple intelligences," popularized by Howard Gardner, and what E. D. Hirsch labels "romantic progressivism," for example. As this article is not meant to focus on philosophy, however, I have kept this digression spare.
72. Daniel M. Koretz, Robert L. Linn, Stephen B. Dunbar, and Lorrie S. Shepard, "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests" (paper presented at the 1991 annual meeting of the AERA, Chicago, April 3–7). See also Robert L. Linn, "Assessments and Accountability," *Education Researcher* (March 2000): 4–16.
73. See a discussion of the phenomenon that includes the physician John Jacob Cannell and many others in full-issue coverage in *Educational Measurement: Issues and Practice* (summer 1988).
74. Test publishers make economic and logistical trade-offs by using convenient samples, such as Chapter 1 students they are already testing to meet Chapter 1 requirements, as norming samples.
75. See Gary W. Phillips and Chester E. Finn Jr., "The Lake Wobegon Effect: A Skeleton in the Testing Closet?" *Educational Measurement: Issues and Practice* (summer 1988): 10–12.
76. Ibid.

77. Shepard, "Will National Tests Improve Student Learning?" 233, 234.
78. Farkas, Johnson, and Duffet, *Different Drummers,* 7; Jean Johnson and John Immerwahr, *First Things First: What Americans Expect from the Public Schools* (New York: Public Agenda, 1994).
79. Shepard, "Will National Tests Improve Student Learning?" 233–34.
80. See, for example, Mary Lee Smith, "Put to the Test: The Effects of External Testing on Teachers," *Educational Researcher* 20, no. 5 (June 1991); "Meanings of Test Preparation," *American Educational Research Journal* 28, no. 3 (fall 1991); "The Role of Testing in Elementary Schools," CSE Technical Report 321, Los Angeles, UCLA, May 1991; and Lorrie Shepard et al. "Effects of High-Stakes Testing on Instruction," paper presented at the Annual Meeting of the AERA, Chicago, April 1991.
81. See, for example, Shepard, "Will National Tests Improve Student Learning?" 233, 234.
82. See, for example, the example on the first two pages of Mary Lee Smith, "Put to the Test: The Effects of External Testing on Teachers," *Educational Researcher* 20, no. 5 (June 1991).
83. Farkas, Johnson, and Duffett, *Different Drummers: How Teachers of Teachers View Public Education,* 12.
84. See OECD, *Education at a Glance,* 1997, p. 200, for the salary figures. See also John H. Bishop: "Impacts of School Organization and Signaling on Incentives to Learn in France, the Netherlands, England, Scotland, and the United States," working paper no. 94-30, Center for Advanced Human Resource Studies, New York State School of Industrial and Labor Relations, Cornell University (Ithaca, N.Y.: December 1994) and "Incentives for Learning: Why American High School Students Compare So Poorly to Their Counterparts Overseas," working paper no. 89-09, Cornell University School of Industrial and Labor Relations (1989) for discussions of the relationship of external tests and teacher status.
85. See Richard P. Phelps, "Benchmarking to the World's Best in Mathematics," *Evaluation Review* 25, no. 4 (August 2001); Linda Ann Bond and Darla A. Cohen, "The Early Impact of Indiana Statewide Testing for Educational Progress on Local Education Agencies," *Advances in Program Evaluation,* in ed. Rita G. O'Sullivan and Robert E. Stake, Vol.1, Part B, 1991, 87–88; or see James Stigler's work comparing time use in U.S., German, and Japanese lower secondary mathematics and science classes, in James W. Stigler and James Hiebert, "Understanding and Improving Classroom Mathematics Instruction: An Overview of the TIMSS Video Study," *Phi Delta Kappan* (Bloomington: Phi

Delta Kappa, September 1997): 14–21.

86. Robert Rothman, "Study Confirms 'Fears' Regarding Commercial Tests," *Education Week* 12, no. 7 (October 21, 1992): 1, 13.
87. Malcolm Gladwell, "NSF Faults Science and Math Testing," *Washington Post,* 16 October 1992, A1, A4.
88. Ibid., 1.
89. Jerry Bobrow, *Cliffs SAT I Preparation Guide* (Lincoln, Neb.: Cliffs Notes, 1994) 63.
90. They are made explicit, however, in the cognitive laboratory testing that some multiple-choice tests undergo when they are developed.
91. Maryellen C. Harman and Claudette Fong-Kong Mungal, *The Influence of Testing on Teaching Math and Science in Grades 4–12: Appendix B: An Analysis of Standardized and Text-Embedded Tests in Mathematics* (Boston: CSTEEP, October 1992): 5.
92. Here's another definition of higher-order thinking that can be used as a point of comparison: "Students engage in purposeful, extended lines of thought during which they: identify the task or problem type; define and clarify essential elements and terms; judge and connect relevant information; and evaluate the adequacy of information and procedures for drawing conclusions and/or solving problems. In addition, students become self-conscious about their thinking and develop self-monitoring problem-solving strategies. Commonly specified higher-order reasoning processes are: cognitive: analyze, compare, infer/interpret, evaluate; metacognitive: plan, monitor, review/revise." (Edys S. Quellmalz, "Needed: Better Methods for Testing Higher-Order Thinking Skills," *Educational Leadership* 43, no. 2 [October 1985]: 30)
93. See George F. Madaus, Mary Maxwell West, Maryellen C. Harmon, Richard G. Lomax, and Katherine A. Viator, *The Influence of Testing on Teaching Math and Science in Grades 4-12: Executive Summary* (Chestnut Hill: CSTEEP, Boston College, October 1992); and West and Viator, *The Influence of Testing on Teaching Math and Science in Grades 4-12: Appendix D: Testing and Teaching in Six Urban Sites.*
94. E. D. Hirsch, of course, makes a more detailed and eloquent argument for the acceptance of both process and content as necessary components of intelligence. See E. D. Hirsch Jr., *The Schools We Need and Why We Don't Have Them* (New York: Doubleday, 1996).
95. Monte Neill, *High Stakes Tests Do Not Improve Student Learning* (Cambridge: FairTest, 1998).

96. FairTest, "Testing Our Children: Introduction" (Cambridge: FairTest, 1998): 2.
97. FairTest, "How the States Scored," and "Vermont," *FairTest Examiner* (summer 1997): 1–3.
98. Neill, *High Stakes Tests Do Not Improve Student Learning.*
99. See, for example, John H. Bishop, "A Steeper, Better Road to Graduation," *Education Next* (winter 2001).
100. John H. Bishop, "Diplomas for Learning, Not Seat Time: The Impacts of New York Regents Examinations," working paper no. 97-31, Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies (1997), 11–17.
101. See, for example, Ludger Woessman, "Why Students in Some Countries Do Better," *Education Next* (summer 2001); Richard P. Phelps, "Benchmarking to the Best in Mathematics: Quality Control in Curriculum and Instruction Among the Top Performers in the TIMSS," *Evaluation Review* 25, no. 4 (August 2001); Debra Viadero, "Assessment Payoff," *Education Week* (September 10, 1997): 32; as well as the work of Robert Costrell, Julian Betts, Thomas Dee, David Grissmer, Anne Danenberg, and others.
102. FairTest, "FairTest Fact Sheet: The SAT" and "SAT, ACT Bias Persist," *FairTest Examiner* (Cambridge: FairTest, fall, 1995).
103. Nancy Cole and Warren Willingham, *Gender and Fair Assessment* (Princeton: ETS, 1997).
104. "ETS Gender Bias Report a 'Smokescreen.'" *FairTest Examiner* (Cambridge: FairTest, fall 1997).
105. Sol Pelavin and Michael Kane, *Changing the Odds* (New York: The College Board, 1990).
106. David W. Murray, "The War on Testing," *Commentary* (September 1998): 34–37; see also Jessica L. Sandham, "Ending SAT May Hurt Minorities, Study Says," *Education Week* (January 14, 1998): 5.
107. "Diversity Takes Back Seat to Standards in New Poll," *Education Daily* (July 30, 1998): 3, 4.
108. Steve Farkus, Jean Johnson, Stephen Immerwahr, and Joanna McHugh, *Time to Move On: African-American and White Parents Set an Agenda for Public Schools* (New York: Public Agenda, 1998): 16, 17.
109. Walter M. Haney, George F. Madaus, and Robert Lyons, *The Fractured Marketplace for Standardized Testing* (Boston: Kluwer, 1993): 119.
110. Ibid., 122.
111. Lawrence O. Picus and Alisha Tralli, *Alternative Assessment Programs: What Are*

the True Costs? CSE Technical Report 441 (Los Angeles: CRESST, February 1998: 47.

112. See Richard P. Phelps, "Estimating the Cost of Standardized Student Testing in the United States," *Journal of Education Finance* 25, no. 3 (winter 2000). See also U.S. General Accounting Office, *Student Testing: Current Extent and Expenditures, with Cost Estimates for a National Examination,* Report GAO/PEMD-93-8 (Washington, D.C.: Author, 1993): 66; and U.S. Department of Education, National Center for Education Statistics, *Digest of Education Statistics 1997,* by Thomas D. Snyder and Charlene M. Hoffman, Washington, D.C.: U.S.GPO, 1997, Table 33.
113. Madaus, "The Effects of Important Tests on Students: Implications for a National Examination System," 227.
114. George F. Madaus and Thomas Kellaghan, "Student Examination Systems in the European Community: Lessons for the United States" (contractor report submitted to the Office of Technology Assessment, June 1991).
115. See Richard P. Phelps, "Are U.S. Students the Most Heavily Tested on Earth?" *Educational Measurement* 15, no. 3 (fall 1996); see also Richard P. Phelps, "Benchmarking to the Best in Mathematics: Quality Control in Curriculum and Instruction Among the Top Performers in the TIMSS," *Evaluation Review* 25, no. 4 (August 2001).
116. Madaus and Kellaghan, "Student Examination Systems in the European Community: Lessons for the United States."
117. See Richard P. Phelps, "Trends in Large-Scale Testing Outside the United States," *Educational Measurement* 19, no. 1 (spring 2000). The study included OECD countries, plus Russia and China.
118. On the latter point, one speaker, Linda Darling-Hammond, then of Columbia University Teachers College, now at Stanford University, said, "In contrast to testing in most other countries, testing in the U.S. is primarily controlled by commercial publishers and nonschool agencies that produce norm-referenced, multiple-choice instruments designed to rank students cheaply and efficiently." (Linda Darling-Hammond, "The Implications of Testing Policy for Quality and Equality," *Phi Delta Kappan* [November 1991]: 220).
119. Robert E. Stake, "The Teacher, Standardized Testing, and Prospects of Revolution," *Phi Delta Kappan* (November 1991): 246.
120. Mary Lee Smith and Claire Rottenberg, "Unintended Consequences of External Testing in Elementary Schools," *Educational Measurement: Issues and Practice* (winter 1991): 10, 11.
121. See Phelps, "The Demand for Standardized Student Testing."

122. Mary Maxwell West and Katherine A. Viator, *Teachers' and Administrators' Views of Mandated Testing Programs* (Boston: CSTEEP, October 1992), Table 3.
123. Ibid., 2.
124. Ibid., 6.
125. Ibid., 39, 40. Other sources of "test invalidity" included "kids are not on grade level," even though a student can be so because he doesn't study or pay attention in class; "kids don't try on tests," even though it can be the fault of the student that he doesn't try; or "tests have weird words, content unfamiliar to the students (language/culture bias)," even though words can be "weird" and "content unfamiliar" because a student doesn't do his reading, study, or pay attention in class.
126. National Association of State Boards of Education, The Full Measure: *Report of the NASBE Study Group on Statewide Assessment Systems* (Alexandria, Va.: Author, 1997); and Millicent Lawton, "State Boards' Leaders Call for Assessments Bearing Consequences," *Education Week on the Web* (October 22, 1997).
127. For an interesting study of the positive opinions of teachers and administrators toward one state test, see Linda Ann Bond and Darla A. Cohen, "The Early Impact of Indiana Statewide Testing for Educational Progress on Local Education Agencies," in *Advances in Program Evaluation* Vol.1, Part B, ed. Rita G. O'Sullivan and Robert E. Stake (1991), 78, 79, 87, 88.
128. Gregory J. Cizek, "The Case Against the SAT," book review, *Educational and Psychological Measurement* 50, no. 3 (autumn 1990): 705.
129. For example, see "State News Roundup," *Education Week on the Web* (June 8, 1994): 1.
130. According to Jeff Moss, the associate school superintendent for the Hoke County, North Carolina, schools, before the accountability reforms, "We had seven levels of instruction for a subject matter—such as seven levels of biology, seven levels of English I—which ranged from remedial to honors or college preparatory. So the teacher expectation was such that if I labeled you a basic student I needed to put you in basic English and not require much from you." See Molpus, "Improving High School Education."
131. For a comprehensive overview of the quality and reliability of teacher evaluations of student achievement, see Richard J. Stiggins and Nancy Faires Conklin, *In Teachers' Hands: Investigating the Practices of Classroom Assessment* (New York: SUNY Press, 1992).
132. Gregory J. Cizek, "Grades: The Final Frontier in Assessment Reform," *NASSP Bulletin* (December 1996).
133. Robert B. Frary et al., "Testing and Grading Practices and Opinions of Secondary

School Teachers of Academic Subjects: Implications for Instruction in Measurement," *Educational Measurement: Issues and Practice* (fall 1998): 23–30.

134. Farkas, Johnson, and Duffett, *Different Drummers: How Teachers of Teachers View Public Education.*
135. Ibid., 10–12.
136. Ibid., 10, 11.
137. Ibid., 11.
138. Ibid., 15, 16.