United States General Accounting Office

GAO

Report to Congressional Requesters

January 1993

STUDENT TESTING

Current Extent and Expenditures, With Cost Estimates for a National Examination

GAO/PEMD-93-8

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-249895

January 13, 1993

The Honorable William D. Ford
Chairman, Committee on Education and Labor
House of Representatives

The Honorable William F. Goodling
Ranking Minority Member, Committee on Education and Labor
House of Representatives

The Honorable Dale E. Kildee
Chairman, Subcommittee on Elementary, Secondary,
and Vocational Education
Committee on Education and Labor
House of Representatives

In 1991, this country began debating in earnest the proposition that the United States adopt a national examination system. It soon became apparent, however, that the debate lacked some key information about the present extent and cost of testing, as well as the likely cost of a national examination system. At your request, we attempted to develop that information and found that students do not seem to be overtested today. Systemwide testing (that is, the testing that is given to all students at any one grade level in a school district) took up about 7 hours for an average student in 1990-91 (half of that time in direct testing and half in related activity) and cost about $15 per student (including the cost of the test and staff time). We estimate that such testing cost about $516 million nationwide in that year, and that a national examination—depending on whether it is based on multiple choice or performance testing—would cost, respectively, about $160 million or about $330 million annually.

We are sending copies of this report to officials at the Department of Education and to others who are interested, and we will make copies available to others upon request. If you have any questions or would like additional information, please call me at 202-275-1854 or Robert L. York, Director of Program Evaluation in Human Services Areas, at 202-275-5885. Other major contributors to the report are listed in appendix V.

Eleanor Chelimsky
Assistant Comptroller General

Executive Summary

Purpose

Recent proposals from the federal executive branch and private groups have drawn unprecedented attention to the idea of a national examination for elementary and secondary students. The House Committee on Education and Labor asked GAO to look at school testing as it exists today, describe its nature, estimate its extent and cost, and assess how a new, national test might affect those factors.

Background

Most of the debate on expanded national testing has centered on major issues of what to test, how to test, and how to use the results. Not much attention has been given to date to the question of how much and what kind of testing there is now. Yet the likely success of future testing may be related to the size, nature, and cost of current efforts about which there exist only wide-ranging, conflicting, and highly uncertain estimates. These range from 30 million to over 127 million standardized tests administered per year, at a cost of from $100 million to $915 million. The congressionally mandated National Council on Education Standards and Testing (NCEST) declined to provide a cost estimate in its report recommending a national testing system, and others' estimates have ranged from a few million dollars a year up to $3 billion.

GAO wanted to obtain valid national data on at least all systemwide tests; that is, those given to all students at any one grade level in a school district. This excludes tests that only selected students take, such as individual teachers' exams, special education diagnostic tests, or college admissions exams. In the fall of 1991, GAO surveyed testing officials in all the state education agencies and in a random sample of U.S. school districts. The survey included questions about each test administered and about the testing officials' views on the balance of costs and benefits in their current testing effort, on trends in the field, and on the idea of a national test. GAO received completed questionnaires from 74 percent of the local districts in the national sample and from 48 of the 50 states. The results are generalizable nationwide.

Results in Brief

In 1990-91, U.S. students do not seem to have been overtested. Systemwide testing took up about 7 hours per year for an average student (half in direct testing and half in related activity) and cost about $15 per student including the cost of the test and staff time. The typical test was the familiar, commercially developed four- or five-subject multiple-choice exam. The less common performance-based tests—in which students write out some answers—cost more (an average of about $20 per student),

but were considered by some testing officials to be an improvement and a preferable direction for further development. GAO estimates the overall cost of systemwide testing in 1990-91 at $516 million.

Three models are commonly discussed for future national testing, including (1) a single national multiple-choice test, (2) a single national performance-based test, and (3) a decentralized system of clusters of states, each cluster using different performance-based tests. GAO estimated that none of these would cost as much as the multi-billion-dollar estimates that some have put forth. The first option would be least expensive ($160 million per year). The third (clusters), the one advocated by NCEST, would likely cost about $330 million per year after about $100 million in start-up development costs, and the costs could be expected to decline over time. Any choice among the three options would involve trade-offs. For example, the least expensive multiple-choice test would be familiar and provide the most comparable data, but would be the most duplicative and might not be as valued by many state and local testing officials. Clusters of performance tests would cost more and would not necessarily be comparable, but may be better linked to local teaching and would be viewed more favorably by many testing officials.

Those officials responding to GAO's survey did not oppose more tests, but expressed concerns over the purpose, quality, and locus of control over the content and administration of further tests. They preferred tests of high technical quality that would be useful for diagnosing problems at the state or local level. However, many respondents expressed opposition to the general idea of a national test.

Principal Findings

The Current Extent and Nature of School Testing

Though the average student spent only 7 hours annually on systemwide testing, GAO found wide variation and totals as high as 30 hours a year. A majority of systemwide testing was state-mandated, with state education agencies developing most of these tests, usually in conjunction with test development contractors. Almost 60 percent of the tests used were commercially available, with achievement tests from three publishers accounting for 43 percent of all systemwide tests. Testing remained traditional in format, with 71 percent of all tests including only multiple-choice questions.

GAO's survey showed that new approaches to testing are finding limited acceptance. By 1990-91, performance-based tests (with the exception of fairly common tests asking for a writing sample) were in use in only seven states or in specialized applications such as readiness tests for very young students. However, these seven states, and several others that have developed high-quality multiple-choice tests, have developed fairly sophisticated testing programs and have gained an expertise in test development that could be useful to the development of a national examination system. Most of these states, moreover, employed local teachers and administrators in test development and scoring and reported that their involvement facilitated acceptance of the test and the alignment of the test to the subject matter that teachers actually teach.

The Current Cost of Testing

The $15 per-student average cost of testing included $4 in purchase costs and over $10 in state and local staff time, but costs varied for different types of tests. In a subset of states where GAO obtained the best comparative data, multiple-choice tests averaged less than half the cost of performance-based tests ($16 versus $33, respectively).

In budgetary terms, testing rarely accounted for more than 1 percent of school district budgets, averaging about one-half of 1 percent. State testing programs averaged less than 2 percent of state education agency budgets. For only three tests in the country did state costs average more than district costs.

The Future Cost and Extent of Testing

GAO estimates that a national test modeled on the common multiple-choice tests, if taken by 10 million students a year, would cost about $160 million; a national performance-based test similar to those now developed in several states would cost $330 million per year, or almost two thirds of the $516 million GAO estimates is now spent on systemwide testing. Start-up development costs could add another $100 million.

But GAO found new costs would vary depending on the plan. Looking at decisions made in school districts that in the past faced a choice between an old test and a new state-mandated test, GAO found that 82 percent dropped the old test when the state's largely duplicated it, but were much more likely to use both if the tests differed in purpose or coverage. If the same pattern held true in response to a national test, a national multiple-choice test would cost the districts only $42 million more and 15 minutes per student in new costs, all from additional testing in 26 percent

of U.S. school districts. The other 74 percent of districts would simply drop a current test, replacing it with the national test. Because many fewer districts use such tests now, a national performance-based test would add more new costs in money and time: $209 million and 30 minutes per student.

Testing Officials' Views on Present and Future Testing

Seventy-five percent of state testing officials and 43 percent of local testing officials considered the net benefits of their present testing programs to be positive, and most believed that these benefits would continue or even increase if more tests were added.

Majorities mentioned performance-based testing as a positive trend and confirmed a trend away from norm-referenced multiple-choice tests toward tests with a higher degree of curriculum alignment. Less than half the states had a curriculum that their districts were obliged to follow, however, while 10 states had unrequired curricula.

The survey revealed significant opposition to the concept of a national examination system. Forty percent of local respondents and 29 percent of state respondents saw no advantages to a national system, and they forecast some disadvantages, particularly a potential for misuse of test results. Thirty-two percent of local respondents and 53 percent of state respondents, however, specifically cited the potential for comparing test scores nationally as an advantage of a national testing system. When asked under what conditions they would decide to use a voluntary national test, they rated most important whether or not the test was of high technical quality, useful to their needs, and not costly to them.

Matters for Congressional Consideration

GAO believes that if a decision is made to implement a national examination system, the Congress may wish to ensure the involvement of local teachers and administrators in test development and scoring and of state testing officials in planning and implementation. This should build support and improve the likelihood of success as state and local educators will probably play a considerable role in the administration of any national test.

If the Congress wishes to encourage the development of a well-accepted and widely used national examination system, it should also consider means for ensuring the technical quality of the tests. Test quality will require an enduring commitment and sufficient resources.

Contents

Figures

Abbreviations

ETS	Educational Testing Service
IQ	Intelligence quotient
NAEP	National Assessment of Educational Progress
NCEST	National Council on Education Standards and Testing
NCTPP	National Commission on Testing and Public Policy
OTA	Office of Technology Assessment
SAT	Stanford Achievement Test or Scholastic Aptitude Test
UCLA	University of California at Los Angeles

Introduction

In the summer of 1991, the country was debating in earnest the proposition that the United States adopt a national examination for elementary and secondary school students. Several proposals with some measure of detail were put forth by various policy-oriented groups.[1] One major stimulus for the proposals was the continuing interest in measuring progress toward the national educational goals that emerged from the September 1989 "education summit" meeting between the state governors and the President at Charlottesville, Va.

National Test Proposals

Since 1990, many in the field have offered a variety of proposals for some type of national testing. These include:

- the American Achievement Tests segment of President Bush's America 2000 education strategy;
- innovative performance-based tests urged by a coalition of university researchers working on the New Standards project;
- a single national multiple-choice test advocated by Educate America, an ad hoc group;
- work-related skill tests recommended by the Secretary of Labor's Commission on Achievement of Necessary Skills; and
- a variety of state tests merged into a national system, proposed by the congressionally mandated National Council on Education Standards and Testing. (The Council's ideas are discussed in more detail in chapter 4.)

The stated principal objective of each of the test proposals was to encourage better teaching and more learning—in short, to raise education standards, and in turn, to improve the economic competitiveness of the nation.

The Debate Over National Testing

Advocates for national testing argued that to compete in a technologically advanced world, American students must achieve higher levels of knowledge and skills. Some argued that the new tests should improve academic achievement by driving instructional practices and curricula to be more focused and challenging than they have been. Some argued, further, that the new examinations should facilitate comparisons across all states and school districts, calling attention to the most successful and deficient education programs.

[1]See, for example, Education Commission of the States, "National Efforts," State Education Leader, 11:1 (spring 1992), pp. 6-10.

Critics agreed that a national test would focus instructional practices and curricula, but thought this would be harmful if the focus was too narrow and some skills and subject matter lost out. Some critics argued, further, that interpretation of academic results shown in the tests should be balanced with information on students and schools because students come to school from widely different backgrounds and attend schools with widely different levels of resources. Other arguments revolved around format and administration. Should test questions have a multiple-choice or performance-based format? Should they be based on a national curriculum, and if so, should the curriculum be developed first or simultaneously? Should a national body develop and administer the test, or should both tasks be left to the states to coordinate by some sort of compact among them?[2]

Early in the debate over national testing, decisionmakers saw that they lacked some key information. What was the current extent and cost (in both time and dollars) of testing in the schools, and how much would a national examination cost? Some opponents of a national exam asserted that American students and teachers were already overburdened with standardized tests.[3] Other opponents asserted that a national examination would be prohibitively expensive; they often based their estimates on the cost of one particular test series.[4]

Estimates of the extent of standardized testing in the United States ranged from 30 million to over 127 million tests administered annually. Similarly, estimates of the current annual cost of standardized testing ranged from $100 million to $915 million.[5] We discuss some estimates of the extent and cost of testing in appendix IV. Estimates of the cost of a new national test varied widely, too. Our survey of the literature revealed seven thoughtful estimates that ranged from several million dollars annually for a multiple-choice test like the Armed Services Vocational Aptitude Battery

[2]Many of these same issues are also addressed in our analysis of setting and measuring standards for student achievement to be used with the National Assessment of Educational Progress, the subject of a forthcoming report.

[3]They often referred to a report of the National Commission on Testing and Public Policy, From Gatekeeper to Gateway: Transforming Testing in America (Boston: Boston College, 1990), pp. 14-18.

[4]This is the Advanced Placement examinations of the Educational Testing Service, which are expensive for several reasons: each subject-area exam is administered separately, to different populations, in highly secure conditions, and test scorers are flown in from many states to central sites.

[5]The low estimates of number of tests and costs come from Douglas J. McRae, "TOPIC: Too Much Testing?," press release, Monterey, Calif.: CTB Macmillan/McGraw-Hill, Nov. 15, 1990; the high estimates are from the National Commission on Testing and Public Policy, From Gatekeeper to Gateway (Boston: 1990).

to $3 billion a year plus $10 billion in development costs for a system of performance-based exams similar to the Advanced Placement series.

Objectives

To obtain more reliable estimates, the House Committee on Education and Labor and its Subcommittee on Elementary, Secondary, and Vocational Education asked us to examine the present extent and cost of testing in the United States. Specifically, we addressed the following questions in our study:

- What is the nature of current standardized school testing, and what is its extent, including tests initiated by local school districts as well as by states?
- What are the costs of these tests?
- How would new national tests affect those factors, and is there any overlap between current assessments and those being proposed?
- How do testing officials view the costs and benefits of present and future testing?

Scope

We restricted the domain of tests to include only "systemwide" tests; that is, those administered to every student, to almost every student, or to a representative sample of all students in at least one grade level in a district or state. Since we intended to use questionnaires as our primary source of data, we realized it was impossible to ask about all tests, or even all standardized tests, because the reporting burden would have been too great and our response rate would have decreased in consequence. The domain of systemwide tests includes all standardized tests except those administered to special populations, such as special education and gifted and talented students; optional tests, such as college entry exams; and many tests used for Chapter 1 evaluation.[6] Thus, the set of systemwide tests seemed the most appropriate for our study, since it consists of the tests most like the national tests proposed for all students. We attempt to account for the extent of other standardized testing in appendix III.

We defined "costs" by its two relevant components. Purchase costs (dollar costs) represent the first cost component of testing—money spent on test-related goods or services purchased at set prices. The test forms and booklets used with a standardized test are purchased from test companies at a contracted price, for example. Likewise, the scoring of

[6]Tests used for federal Chapter 1 program evaluation would only be included in our survey data if the tests were administered at all schools in a school district.

machine-readable forms is a service purchased at a contracted price. Time spent by education personnel represents the second cost component of testing; that is, the amount of time spent in all the test-related activities of developing, administering, preparing for, taking, grading, and interpreting tests by all the parties involved—teachers, administrators, clerical staff, and others. Some of this time is explicitly paid for, and we gathered information on its cost. This cost can also be indirect, or in-kind, if it is not paid for.

In general, any one test should not necessarily be preferred over another simply because it is less expensive, as it may also be less beneficial. We did not attempt to make a quantitative estimate of tests' benefits, as they do not lend themselves to precise measurement. But we did ask knowledgeable officials to give us their assessment of the benefits of testing relative to the costs. We did not limit the type of test or test format we asked about. Many advocates of a national examination system proposed that it employ some of the newer testing techniques, such as performance-based formats (in which students must write, perform a laboratory experiment, or in some other way do more than simply answer a multiple-choice question), since many experts consider such tests of higher quality. So we also sought information pertaining to their cost and extent of use. We buttressed our cost estimates for performance testing with figures obtained from our interviews with education officials in two Canadian provinces that employ performance tests.[7] With adequate data on the cost and extent of most types of current tests, we also planned to estimate both the expense of any proposed national test that would be similar to current tests and its overlap with current testing.

Methodology

Surveys on State and Local Testing

We gathered the primary data to answer the four evaluation questions through surveys of state testing officials and local school district administrators. (Table 1.1 shows the data sources we used for each of the four evaluation questions.)

[7]These interviews were conducted as part of a related study. A detailed discussion of Canadian provinces' experience with school testing will appear in a forthcoming report.

Table 1.1: Sources We Used to Answer Evaluation Questions

Question	Data source
Current nature and extent of school testing	Surveys of all states and a sample of districts; data on each test
Cost of testing	Surveys, data on each test, plus interviews with officials of testing firms, interviews with Canadian officials
Potential effects of new national tests on nature, extent, and cost of testing, and overlap between current and proposed tests	Surveys, case studies of 50 districts, our analysis of national test proposals, plus interviews with officials of testing firms
Testing officials' views on the costs and benefits of current and future testing	Surveys

During July and August 1991, 10 state and local testing officials from four states reviewed early versions of our questionnaires and suggested revisions.[8] We then pretested the revised versions with four local school district officials and one state testing director in Maryland and Virginia. The four local districts represented small, large, and very large student populations and urban, suburban, and rural areas.[9]

We designed two questionnaires for our state or local respondents. The first requested general information about the state or district and the respondents' views on general testing issues. The second requested information about each systemwide test, particularly detailed information on time and dollar expenditures. Respondents were to fill out a separate questionnaire for each test. In hopes of increasing the willingness of officials to respond, with the agreement of the congressional requesters we promised not to identify any specific state or school district.

In September 1991, we sent the questionnaires to all 50 state testing directors and to 663 local public school administrators (director of testing or superintendent) in school districts containing more than 50 students. To achieve a high response rate, we sent the survey twice, if necessary, and then sent two postcard reminders. We telephoned many of the respondents who returned incomplete questionnaires in order to fill in missing information.

Of the 663 local districts that received questionnaires, 15 either had fewer than 50 students, were defunct (usually through merging with another district), or were unable to respond, giving us a total sample size of 648

[8]Those four states were California, Maryland, North Carolina, and Virginia.

[9]We classified district size as small (from 51 to 3,500 students), large (from 3,500 to 35,000 students), and very large (over 35,000 students).

local school districts. Of those districts, 500 formed a nationally representative sample we had designed to produce generalizable estimates for the United States.[10] We received 368 completed questionnaires from this group, for a 74-percent response rate. We received completed questionnaires from 48 of the 50 states. The two remaining states did not administer statewide tests in 1990-91. Appendix I contains our analysis of the sample survey response rates among different respondent groups.

We searched for published information on variations in testing programs among local school districts so that we could target our surveys to cover the different situations. We found no useful information aside from the types of state-mandated tests. We therefore designed our stratified sample using some school district characteristics that were available to us—district size, metropolitan status (urban, suburban, or rural), and type of state test—that we thought to be related to the level and cost of testing.[11]

In addition to surveying the 500 school districts that formed our national sample, we oversampled in certain states that were using performance-based formats in state-specific and state-managed tests.[12] We attempted to get more responses from district officials in these states because there are few data available elsewhere on the implementation of these techniques and their administration costs. Oversampling allowed for more precise estimates of the cost of performance-based tests.

In summary, the surveys provided direct answers to the first two questions concerning the nature, extent, and costs of current testing programs. We also gathered data on test development and its costs from state testing officials and representatives of commercial testing firms. The surveys were also useful in answering the third question—concerning the overlap between current and proposed tests—by providing information on school district reactions in the past when they faced new state test mandates and had to choose between simply adding another test to their programs or dropping a current test in favor of the new state test. The surveys also

[10]The remainder of the 648 local school districts (148 districts were not included in the nationally representative sample) were used to oversample in certain categories of school districts, including those in states that we knew employed statewide performance-based tests.

[11]We classified state tests into four types: no state test, only norm-referenced multiple-choice test(s), at least one criterion-referenced multiple-choice test, and at least one performance-based test.

[12]These states included Arizona, Connecticut, Maine, Maryland, Massachusetts, New York, and Vermont. We oversampled districts in Arizona and Vermont to obtain data on the implementation of portfolio assessments. That effort proved unsuccessful, as most of the respondents in those states did not consider portfolios to be "tests" and thus did not complete surveys about this activity.

provided direct answers to the fourth question regarding testing officials' views on the costs and benefits of present and future testing.

Case Studies of the Effect of Testing Mandates

To find further information on the third question, concerning the possible effects of new national tests on the nature, extent, and cost of current testing programs, we made a separate study of past instances where states mandated that their districts administer new tests. Given new national tests, school districts would face the choice of adding another test to their lineup, discarding an existing test in favor of a national test, or rejecting the national test in favor of the existing tests.

Before the late 1970s, few states mandated statewide tests. By the end of the 1980s, the situation reversed so that only a few states did not do so. From our survey responses, we identified about 200 school districts that had been administering tests of similar subjects and purposes at the time their states imposed statewide tests. Some of these districts kept their old tests; some discarded them. We asked all of them in the main survey how similar in purpose or content the new test was to the old, and specifically, if they dropped the old test and, if so, why. We interviewed officials by telephone in a systematic sample of 50 of these districts to learn more about how they made their decisions and how the new tests affected the level and costs of their testing programs.

Study Strengths and Limitations

The most important strengths of our study are four. First, it is unique, since no other up-to-date information on current testing is available, and new test costs have typically been crudely estimated. Second, our findings are comprehensive, covering the entire country, close to the full population of states and a representative sample of local school districts. Third, with the stratified sample design, we can make stronger estimates of the extent and cost of testing in the United States than we otherwise could. Fourth, we oversampled in certain groups that otherwise would not have been well represented, such as very large school districts and districts in states with statewide performance-based tests.

Of course, the study has some limitations, too. It covers 1 year only—the 1990-91 school year—and we only surveyed public schools. We did not gather information on all testing, or even on all standardized testing, nor did we make first-hand observations to check survey answers, and so any effort on our part to portray the total testing burden on elementary and secondary students (or its cost) from the estimates our respondents gave

can only be roughly approximate. We were not able to collect much data pertaining to assessment methods that testing officials often do not consider to be tests, the most prominent example being student portfolios.

Some useful data were beyond our ability to collect, such as the views of parents, advocacy groups, and students on the costs, burdens, and benefits of testing or the merits of expanded national tests. Similarly, we could not gather first-hand data on key topics such as how tests are used or how they affect instruction; the views of our survey respondents give only some aspects of these matters, and from a particular point of view. Finally, because we asked respondents in fall 1991 their general views on national tests, the answers do not reflect the specifics of any proposals made since then.

Organization of the Report

Chapter 2 addresses the first of our four questions, with estimates of the current nature and extent of systemwide testing in the United States and the relative prominence of different types of tests. Chapter 3 answers the second question, on the current costs of systemwide testing in the United States, in time and in dollars, and for different types of tests. Chapter 4 responds to the third question with information on the possible effects on district testing programs of adding a new test and the conditions under which districts would adopt new tests. Chapter 5 answers the fourth question, providing the views of our respondents on the benefits of their testing programs, and it includes additional views on trends in testing, a national examination system, and current levels and costs of testing programs. Chapter 6 proposes some matters for congressional consideration raised by this study.

The Current Extent and Nature of School Testing

Many claims have been made concerning the number of tests students in the United States are required to take and the amount of time they spend taking them. On the one hand, some state-level education reformers of the past 2 decades thought that there was too little testing to ensure accountability for education expenditures, and they successfully urged expanded statewide testing. On the other hand, some testing critics have asserted that U.S. students take the most tests of any students. In the discussions on national testing, observers reacted to new testing proposals, in part, based on their view of the extent of testing at the time.

This chapter presents our national survey data, allowing us to describe the extent and nature of systemwide school testing for the year 1990-91, including the amount of time students spent in testing, the types of tests they took, who designed these tests, what designs they used, and for which purposes they intended to use them. We also present information on how testing officials see tests linked to curriculum.

Amount of Time Spent in Testing

For the systemwide tests we looked at, the burden of testing on U.S. students was modest in 1990-91. On average, students spent less than 4 hours each taking systemwide tests, or less than one-half of 1 percent of a student's school year.[1] Counting all the time devoted to test-related activities, such as learning test-taking skills or listening to test instructions or results, the mean time burden still averaged less than 7 hours for the year (with the median at less than 6 hours).

There was a wide range to this time burden, however. One school district gave no systemwide tests at all in 1990-91; six districts in our sample administered 10 or more. Eighty-five percent of school districts administered one to three tests (the mean was 2.5; the median, 2 tests a year). Five states required no statewide tests, while one of them required four. The mean number of tests among all states was 1.7; the median was 1 test.

At the high end of the range of testing effort, we found several districts administered over 27 hours of systemwide tests in 1990-91.[2] Counting student time devoted to all test-related activities (including preparation,

[1]The exact number is 3.4 hours. This statistic represents the mean for all U.S. students; the median was 3 hours per student.

[2]This is the sum of the hours required to administer all of the district's tests. No individual student in 1 year would have taken all of them, owing to the common practice (as we describe) of scattering systemwide tests across the grades.

listening to results, and the like), several districts claimed over 100 hours. When we divided each district's total hours by the number of students in the district, we found districts have made a wide range of choices of how much time to devote to testing: from zero to over 13 hours for the average student just writing exams in 1990-91 and from zero to over 40 hours altogether (or a full week of 6-hour school days) in test-related activity.

The most time-consuming test was a certain state test that covered four subject areas. Passing this test was required for graduation. Because the stakes were high, students took the test during a 3-day period with virtually no time constraint on each section of the test. (The official who completed our survey estimated 18 hours for the full test.) Also because the stakes were high, some districts in the state spent a considerable amount of time in test preparation activities.[3] Few tests with more conventional time limits occupied more than 10 hours total test-taking time.

We found more hours of systemwide testing in the school districts with more experience in testing, that have a relatively high level of poverty, that administer high-stakes tests, and that are located in Northeastern or Southern states.[4] Northeastern states' testing programs commonly used longer, performance-based tests that contributed to more hours of testing there. High-stakes testing in Southern statewide testing programs contributed to more hours of test-related activity there, though not more hours of test writing time. On average, high-stakes tests required 43 percent more time in test-related activities other than taking the test, mostly in test preparation activities.

We found fewer hours of systemwide testing in school districts with higher professional salaries and in Western states.[5] Metropolitan location, district size, and the presence of bilingual students seemed to have little clear relationship to the amount of district testing one way or another.

[3]These were defined in our survey as "minutes of instruction in test-taking skills, of taking practice tests, or in motivational activities geared to this test."

[4]Poverty was measured by the proportion of students receiving free or reduced-price lunches. High-stakes tests are those used to determine promotion, retention, or graduation. "High-stakes" tests and tests used for "student-level accountability" are considered synonymous.

[5]Salaries and expenditures represent both the wealth and the cost of living in a region. This includes, generally, the Plains, the Rocky Mountain states, and the Pacific states.

Types of Tests

One way to categorize tests divides them according to the main purpose intended by the test makers. Classified this way, 81 percent of all systemwide tests taken in 1990-91 were achievement tests, those that attempt to measure a student's accumulated knowledge or skill. Most of the achievement tests were commercially available; many of them were state-specific (i.e., designed or adapted to match a state's curriculum). Examples of the more widely used commercial achievement tests included the Iowa Test of Basic Skills, the Comprehensive Test of Basic Skills, the Stanford Achievement Test (SAT), the California Achievement Test, and the Metropolitan Achievement Test.

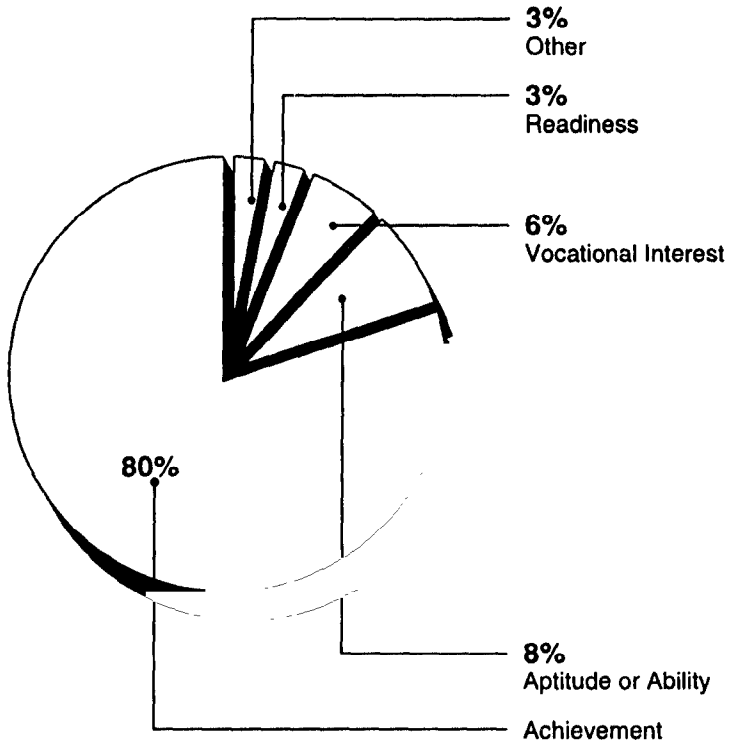
Another 8 percent of systemwide tests were designed to measure aptitude or ability (i.e., future performance). IQ tests fall in this category. Examples of commercially available aptitude tests include the Otis-Lennon School Ability Test, the Cognitive Abilities Test, and the Test of Cognitive Skills. Six percent of systemwide tests were designed to measure vocational interests in order to help students with career planning. Another 3 percent of tests taken in 1990-91 were developed to measure "readiness." Readiness tests are normally given to kindergarten or primary school students. Figure 2.1 summarizes test types according to their main purpose.

Another way to categorize tests divides them according to the subject areas covered; of course, any one test could address from one to several subjects. Systemwide tests in 1990-91 mostly covered school achievement in five core subjects: math, reading, grammar, science, and history or social science.[6] Our respondents claimed that 25 percent of systemwide tests addressed aptitude or ability and 10 percent addressed readiness, though for that to be true, some school districts must have been using tests that were designed to measure achievement as an indicator of aptitude, ability, or readiness. As the previous section explained, only 8 percent of tests were designed to measure aptitude and only 3 percent of tests were designed to measure readiness.

Our respondents also told us that 36 percent of tests addressed writing, 12 percent "critical thinking," 7 percent civics or citizenship, 6 percent vocational interests, and 1 percent attitudes. Notable in their absence (at less than 1 percent of districts) were tests that included foreign language or art. Seldom do all students in a district take art or any single foreign language, so these subjects tend not to be tested systemwide.

[6]Seventy-eight percent of tests addressed math knowledge or skills, 70 percent reading, 49 percent grammar, 44 percent science, and 42 percent history or social science. Percentages do not total 100 because many tests cover multiple subject areas.

Figure 2.1: Types of Tests[a]



[a]The sum of the percentages exceeds 100 because of rounding.

On the question of the grade-level at which tests were given, we found, first, that local tests were distributed fairly evenly over all the elementary and secondary grade levels, with some drop-off at 12th grade. The distribution of state tests over the grade levels was more uneven. More state tests were given in grades 3, 4, 6, 8, and 11. Very few state tests were administered to 1st, 2nd, or 12th graders.

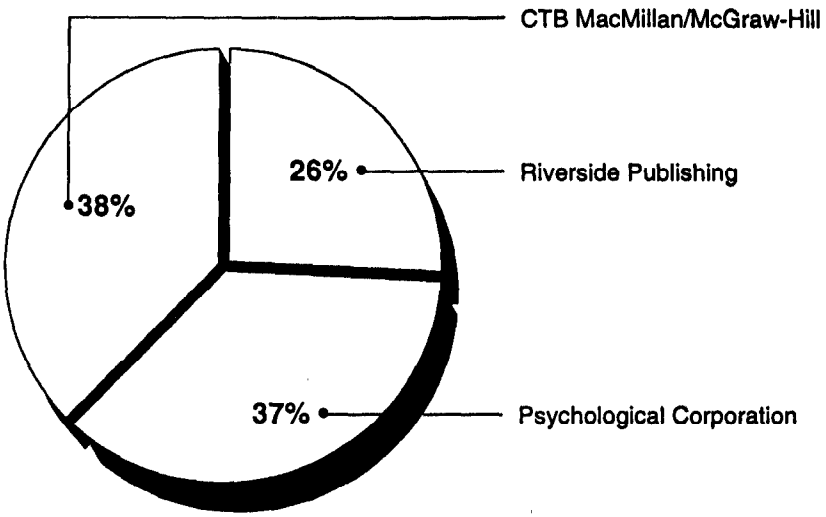
Sources of Tests

State education agencies strongly influence local school district testing programs. States mandated just over half of all the systemwide tests administered in U.S. school districts. State education agencies developed most, but not all, of these state-mandated tests, usually in conjunction with test development contractors. Half the state education agencies required that their local districts administer specific commercially developed tests. In four cases, states modified these tests to match state curriculum standards. Another four state education agencies required only that their

local districts administer some commercially developed test from an approved list.

Thus, directly or through state mandates, commercial test publishers are also a force shaping school district testing programs. Almost 60 percent of the systemwide tests reported to us were commercially developed. In fact, achievement tests produced by the three largest commercial test publishers comprised 43 percent of all tests.[7] Figure 2.2 summarizes these data on sources of tests.

Figure 2.2: Sources of Commercially Developed Tests[a]



[a]The sum of percentages exceeds 100 because of rounding. Tests include four categories: achievement, aptitude, readiness, and vocational interest. CTB MacMillan/McGraw-Hill published the Comprehensive Test of Basic Skills, California Achievement Test, Science Research Associates tests, Test of Cognitive Skills, and Kuder Occupational Interest Survey. The Psychological Corporation published the Stanford Achievement Test, Metropolitan Achievement Test, Otis-Lennon School Abilities Test, Differential Aptitude Test, and Ohio Vocational Interest Survey. Riverside Publishing published the Iowa Test of Basic Skills, Iowa Test of Educational Development, Tests of Achievement and Proficiency, Cognitive Abilities Test, and the 3-R's Test.

Test Design

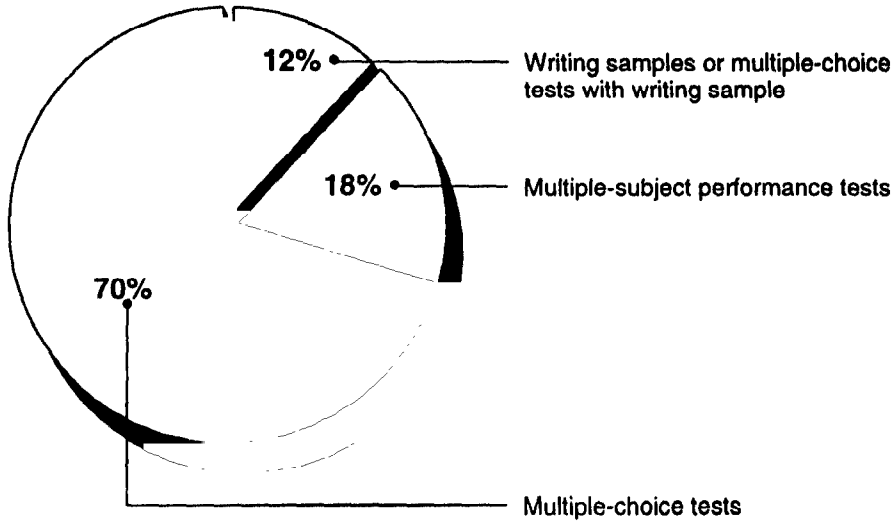
When we asked about the technical design of current tests, we found most were quite traditional, despite lively debate in recent years about needed improvements. That is, most systemwide tests were designed to show how a student performs in relation to the norm or average of all others

[7]Some of the tests were customized to state or local specifications. The three publishers are: CTB MacMillan/McGraw-Hill, the Psychological Corporation/Harcourt-Brace-Jovanovich, and Riverside Publishing/Houghton-Mifflin.

(norm-referenced) and to measure knowledge by asking students to choose one answer among several choices for each of a battery of questions (multiple-choice format). The alternatives are to measure a student against some standards or criteria external to the group (criterion-referenced) or to examine more types of skills and learning by calling for short written answers, essays, or other more creative activities (performance-based format).

In the critical appraisals of a majority of testing experts and the larger community of education professionals, criterion-referenced and performance-based tests are more popular than the traditional norm-referenced and multiple-choice tests. Responses to the opinion questions of our survey affirm this (see chapter 5). Yet, testing practice lags behind these preferences. We found that 71 percent of the tests administered last year were norm-referenced, reflecting the dominance of national commercially developed tests. And 71 percent of the tests were formatted exclusively with multiple-choice responses. Thirty percent of the tests did contain some performance element, but 40 percent of them were writing samples alone or test batteries that included a writing sample but were multiple-choice tests. Only 18 percent of all tests asked students to perform in more than one subject area using performance formats. Figure 2.3 summarizes the test design features we found.

Figure 2.3: Test Design Features[a]



[a]The sum of the percentages exceeds 100 because of rounding.

Writing samples, reading comprehension and response exercises, and math or science problem-solving predominated among the performance-based test formats. Less frequently, we found some use of other types of performance formats, such as science laboratory work, group work, or skills observations. But laboratory work and group work comprised only 4 percent of all performance formats used in 1990-91. Skills observations comprised a larger percentage—12 percent—but largely because of their use in readiness tests. Thus, performance formats remain dominated by the more traditional paper-and-pencil essay questions.

States and school districts, rather than the testing industry, seem to have managed most of this type of testing up to now. Table 2.1 shows that performance-based tests in 1990-91 tended more often to be state-mandated and to be much more often developed by or for a state or school district than tests in general.

Table 2.1: Comparison of Performance-Based Tests With All Tests

Test characteristic	Performance-based tests	All tests
State-mandated	88%	58%
Developed by or for school district	14	7
Developed by or for state	76	35
Commercially developed	9	55
Grades K-8	77	82
Grades 9-12	55	56

A few more states are now trying criterion-referenced, performance-based statewide tests, and all of the three largest test publishers expect to have their major achievement exams available with performance formats within the next 2 years. The costs of performance-based tests are discussed in chapter 3.

Purposes of Tests

Another debate over testing concerns the stakes that should ride on the results, with higher stakes thought by some to strengthen teacher and student motivation, but by others to divert too much time from regular classwork and even to prompt cheating by teachers and students. In fact, most districts reported low stakes. Districts gave tests because their states required them and because they believed tests offered useful information on the students, schools, or curriculum. Thus, we found that local districts were least likely to report using tests for student or school accountability or for student placement. Twenty-four percent were reported formally used for student accountability and, therefore, as "high-stakes" tests. (See glossary.) For over half of the tests administered, the respondents rated student or school accountability measures of little or no importance or not applicable.

At the state level, however, district or state accountability was a vivid purpose for testing—though not student or school accountability. States reported a clear purpose in making test results public to encourage voters or school boards to instigate needed systemwide changes. As was true at the district level, though, state education agencies most commonly administered statewide tests for purposes of evaluation and diagnosis. The least popular uses of statewide tests involved state-level management (planning, tracking, or resource allocation) or grouping and placement of individual students.

Relationship of Tests to Curriculum

Whether students have had enough opportunity to learn the material on tests is another continuing issue in testing policy debates. The match of what is tested with what is taught—or required to be taught—is sometimes referred to as the alignment of the two, and state curriculum requirements are one means toward that end. Despite considerable discussion of the need for standards prescribing course content, not all states had a statewide curriculum in 1990-91, and not all of those that did required their local districts to follow it. At least 17 states had no curriculum and at least 10 others had curricula their local districts were not obliged to follow. Only 14 states both required that local districts follow a state curriculum and administered a statewide test. For 65 percent of the statewide tests in those states with a curriculum, officials told us they believed the tests were largely or perfectly aligned with the curriculum, and for another 30 percent, officials believed the tests were moderately aligned.

Local district respondents reported that 37 percent of the districtwide tests in use in 1990-91 had caused some curricular realignment, 27 percent to a moderate or large extent. The influence of tests on curriculum was judged positively, by and large. Where local officials reported shifts in curriculum in response to tests, about two-thirds thought that the realignment had strengthened learning in their district, while only 2 percent thought that it had weakened learning.

The issue of alignment raises the question of alignment to what, especially if local teaching and curriculum do not match the breadth or depth of content national experts recommend in an area. As shown in chapter 5, some state testing officials told us they prefer tests geared to their curricula, though that may to some degree be at odds with the current pressure for schools to adopt national standards and be tested against them.

Trends in State Testing Programs

As was mentioned in the previous chapter, few states mandated statewide tests before the late 1970s, but by the end of the 1980s, few did not do so. Many of the first statewide tests, arising from the "back-to-basics" emphasis of the period, were meant to measure "minimum competency." They tested only the major subjects and sometimes just reading, writing, or math. More often than not, states merely purchased, or required that their local districts purchase, commercial norm-referenced tests. Partly in reaction to perceived shortcomings in this method of assessment, state education officials argued for different testing programs. In many states,

they were allowed to design testing programs that largely matched their desires.

Greater control over student testing by state education officials has fostered several trends. They include: more involvement in test development by state and local education officials; more criterion-referenced testing and less norm-referenced testing; more performance-based formats; teacher involvement in test development and scoring; test development procedures that include consensus-building among most interested groups; collecting and releasing social and economic indicators, along with test results, to describe school district or performance; and statewide testing programs incorporating more than one test.

Local teacher and administrator involvement in test development and scoring has generally worked to the satisfaction of all parties in many states and Canadian provinces. Moreover, survey respondents in states with criterion-referenced performance-based tests—which provide an opportunity for teacher involvement in development and scoring—usually cited teacher involvement as one of the major strengths of their testing programs. Not all teachers and administrators need to be involved, just enough, on a rotating basis, to give local education professionals a sense that their group is influential in the process.

Some developments, occurring in too few states and too recently to be called trends, point to ways in which state testing programs might be expanded. Two states are attempting to develop programs that are relatively comprehensive in subject matter, including tests in art, music, many vocational education subjects, and more. Five states are in the early stages of development for statewide end-of-course tests. Two states already administer statewide achievement tests for advanced high school subjects, and other states may join in that effort.

Thus, state education agencies are actively involved in testing and are getting more so. Testing activity has been stalled some in several states owing to current poor state fiscal conditions. But though some states have been forced to skip a year or stretch out development schedules, few have given up on statewide tests without replacing them with other tests.

Summary

Our survey results suggest that testing officials did not ascribe much importance to tests as student-accountability measures but that

one-quarter of all tests were, nonetheless, high-stakes tests. And with the exception of writing samples, despite all the enthusiasm surrounding criterion-referenced, performance-based testing, by 1991 it was still primarily implemented in the seven states with statewide performance tests and could otherwise be found mostly in early grades' school-readiness tests.

In the main, students in 1990-91 were tested in four or five subject areas, using commercially developed, multiple-choice, norm-referenced, and state-mandated tests. But if state testing officials have their way, more tests in the future will be performance-based, criterion-referenced, and at least partly developed by state and local officials.

At least on average, and considering only systemwide tests, students do not seem to have been overly tested. The average student spent less than 4 hours in the year taking exams. Thus, an argument that a national examination system should be opposed on those grounds demands other evidence than what we found. However, there was a range to the amount of testing from district to district. Some districts tested quite a lot, and some not at all.

The Current Cost of Testing

Of all the unsettled issues in the debate over a national examination, none has provoked such a diverse set of claims as its estimated cost. These have ranged widely—from a few million dollars to several billion dollars a year. The costs of current tests arouse controversy, too, and are not always known precisely. This is true even for tests that are commercially developed and sold at a fixed price, for while the testing firms know their costs, variations in use by the purchasing school districts affect the overall costs in ways that have not been thoroughly documented.

This chapter answers part of the second evaluation question with the results of our surveys on the costs of particular types of tests and on the aggregate cost of testing in the United States. Both allow us to make reasonable estimates of the potential cost of different kinds of national examination systems, a task undertaken in chapter 4. The first part of this chapter discusses the major components that make up the cost of a test, which partly explain why cost estimates can vary so much when taking only some of these components into account. We then present our cost estimates for systemwide testing in the United States, for particular types of tests, and for test development. We also investigate the presence of economies in large-scale testing.

Dollar and Time Costs of Tests

Cost estimates can be thoughtful and accurate and still vary widely, since a test's cost has many components, not all of which are always included in estimates. Some are obvious. The length of the test is one component, for example, and longer tests tend to be more expensive to develop, administer, score, and report than shorter tests when all other factors are equal. Some components are not so obvious. The time taken from a teacher's schedule to administer a test, for example, is often neglected in cost calculations. Test development costs, likewise, often get left out.

Since we asked about all costs in our surveys, we can use the responses to estimate all costs involved in administering systemwide tests in U.S. school districts in the year 1990-91.[1] Our respondents accounted for testing costs in two ways: by listing the dollars they paid out for tests or test-related services or supplies and by estimating the personnel hours

[1]We did not ask about costs included in general school district or state agency overhead expenses, such as the costs of building space used for tests. Such indirect costs would have been difficult for respondents to allocate consistently.

devoted to testing and to test-related activities.[2] For state-mandated tests, we incorporated costs from both the state and the local district level. We calculated the time cost by multiplying the number of hours spent on test-related activity by the hourly employee salary.[3] Adding the time costs to the other costs gave us the total for each test.

Without exception, every test incurred some expenditure of personnel time. School personnel (usually teachers) administered almost all the tests taken by their students. School districts also expended cash when they purchased tests from commercial test publishers. In many cases, however, school districts paid nothing—in cash—for tests: states that developed their own tests commonly did not charge the districts for them. Occasionally, tests were also provided free when a school district served as a pilot for a new test or when it used the Armed Services Vocational Aptitude Battery.

In the year 1990-91, state and local educational agencies paid an average of about $4 for each individual student test administration. At the same time, they devoted slightly over $10 worth of state and local education personnel time for each individual student test administration (that amounts to about 35 minutes of personnel time per student test, or about 620 hours per district test). So each time a student took a test, it cost about $15. On average, each school district expended about 1,500 personnel hours last year on systemwide testing and spent, in dollars and time, about $34,500.[4] In budget terms, testing did not often account for more than 1 percent of school district budgets, averaging about one-half of 1 percent. State programs averaged less than 2 percent of state education agency budgets.

We found wide variation in these figures (from less than $1 to over $90 per student test), so we looked for explanations of those variations. First, the type of test influences costs. Multiple-choice-only tests averaged around $14, while tests with at least some performance component averaged about $20. Second, different districts face different situations that seem to

[2]Respondents were asked to account for the amount of personnel time devoted to: developing the test; preparing students to take the test; getting trained to administer or score the test; training others to administer or score the test; administering the test; collecting, sorting, and mailing completed tests; scoring the tests; and analyzing and reporting the results.

[3]We asked for the time spent on testing by three levels of staff: managerial, nonmanagerial professional, and clerical. We also asked each state or district to give the average salary of each of the three levels, which we then used to calculate the dollar costs of the time spent.

[4]These particular district averages for personnel hours expended and cost do not include any state time and cost figures.

be systematically related to the costs they incur for testing. These factors are summarized in table 3.1.

Some cost variations reflect characteristics of the student body, such as more low-income or non-English-speaking students. Still others reflect state mandates. The choice to use more of the more expensive performance tests carries obvious cost consequences. Northeastern and Southern states may have higher testing costs because they administer the more expensive performance-based tests (in the Northeast) and high-stakes tests with higher levels of test security. In situations where we found a district spending less per student on testing, we also found such features as larger size of district, more grade levels tested, and more experience with the chosen tests, as well as more testing overall. We also found more use of high-stakes tests associated with lower costs.[5]

Table 3.1: Factors Related to Higher and Lower Testing Costs per Student[a]

Testing costs	Contributing factors
Higher	Higher number of performance tests
	Higher proportion of low-income students
	Higher proportion of bilingual students
	State mandates to test
	Northeastern location
	Southern location
Lower	Higher number of tests administered
	Higher number of grade levels tested
	Higher number of years of experience with a test
	Higher number of high-stakes tests
	Larger district size

[a]As measured by cost per student test hour.

Differences in State and Local Costs

As might be expected, local school districts and state agencies differed in the contribution of different kinds of staff and activities to the overall cost figures. For example, in the local districts, teachers and specialists contributed 86 percent of the time spent in test-related activity, and administrators and clerical employees only 12 percent. In contrast, state officials responding to our survey reported that administrative and clerical

[5]High-stakes tests influenced overall testing costs in two different ways. They tended to consume more personnel time in test preparation activities, but these increased costs were more than offset by cost decreases associated with the fact that these tended more often to be multiple-choice tests.

employees contributed about 41 percent of the time spent in test-related activity and nonmanagerial professionals contributed 59 percent.

Concerning test-related activities, states tend to develop rather than administer tests, while districts show the opposite pattern: much administrative expense but few development costs. Thus at the district level, 39 percent of time was devoted to administering tests; 28 percent to preparing students; 18 percent to collecting, scoring, and analyzing the tests; and 15 percent to other test-related activities. At the state level, 36 percent of time was devoted to test development; 10 percent to training; 37 percent to scoring, collecting, mailing, and analyzing; and 17 percent to other activities. Only 9 percent of state-level time was devoted to test administration, and only 2 percent of district-level time was devoted to test development.

For only three tests in the United States did state costs average more than district costs. Even in those states administering their own state-developed, full-battery (that is, three or more core subject areas) performance-based tests in 1990-91—probably the most expensive possible situation for a state—district costs exceeded state costs. On average, the state assumed only 25 percent of the costs of tests in which states were involved. Even with tests that state agencies themselves developed, printed, distributed, scored, analyzed, and provided to the districts without charge, the bulk of the costs fell at the local level. The result reflected the fact that personnel time devoted to test administration always comprised the majority of the costs, and these were, of course, costs only to the local school districts.

Multiple-Choice and Performance-Based Test Costs

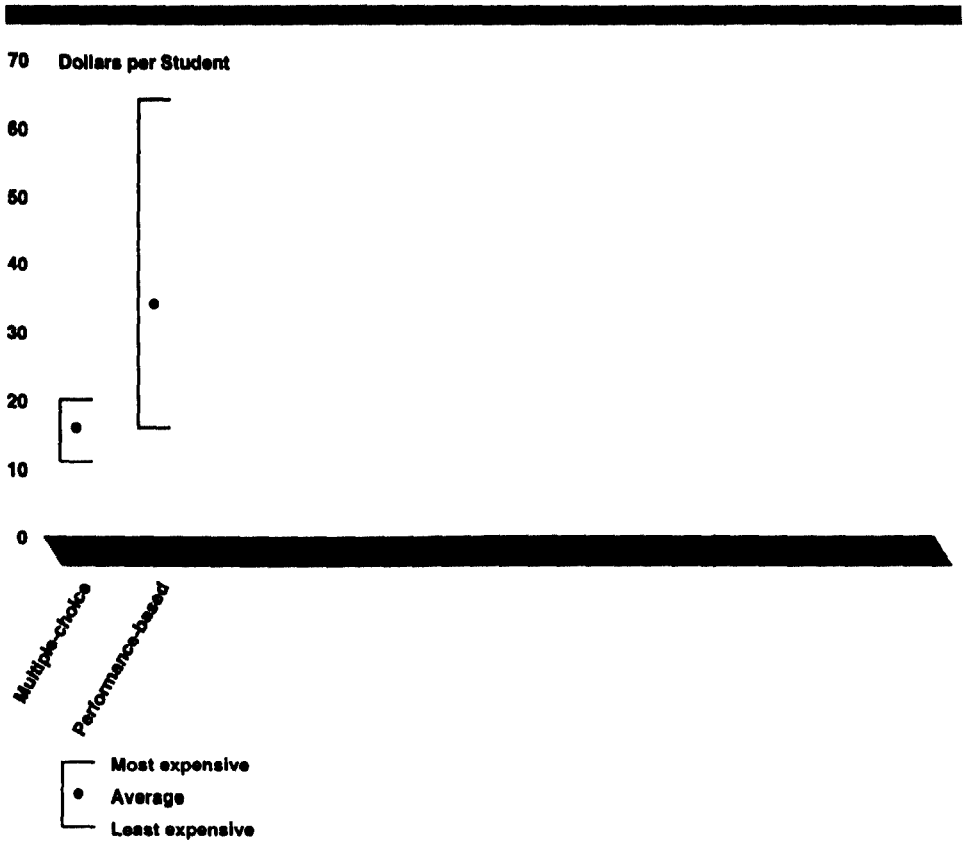
As we learned in the opinion section of our survey (reported in chapter 5), and as widespread discussion of desirable improvements in testing shows, there are currently both great hopes and large unknowns about new methods of testing that go beyond asking students to choose from among several answers. These performance-based tests are known to be more expensive than multiple-choice tests as a general rule, but how much more expensive? Accurate estimates require clear distinctions among different definitions of a performance test. Many have some multiple-choice items and may have only one or several performance components. Formats can vary widely in type and expense, and as a result, performance-based tests can vary widely in cost.

Our large survey sample and oversampling in states with state performance-based tests allowed us to obtain a good comparison of the costs of the two types of tests by looking at school districts in states where both were administered. Thus, we could hold constant, or remove the confusing effect of, many factors by examining costs of two different kinds of tests in the same district for the same student population, and all as reported by a single person completing our survey. And where school districts generally administer both kinds of tests, the performance-based tests are likely to be clearly different from the multiple-choice tests, different enough to justify using both.

In the six states where school districts used both state-developed performance-based tests and commercially developed multiple-choice tests, we found the performance-based tests were typically almost twice as expensive. As shown in figure 3.1, the multiple-choice tests averaged $16 per student (ranging from $11 to $20), while the performance-based tests averaged $33 (with a range from $16 to $64).[6]

[6]Strictly speaking, these cost figures may underestimate the cost of pure performance-based tests. These six states reported to us on a total of 11 performance-based tests (two states used 2 each and another state used 4). Of the 11 tests, only 1 was formatted exclusively with performance questions. All of the others had some multiple-choice questions. All of the tests, however, employed performance formats in more than one subject. That distinguishes these from other state tests with performance-based formats in writing but multiple-choice formats in all other subject areas. The percentage of test time devoted to performance-based questions among these tests ranged from 20 to 100, with a mean of 46 percent.

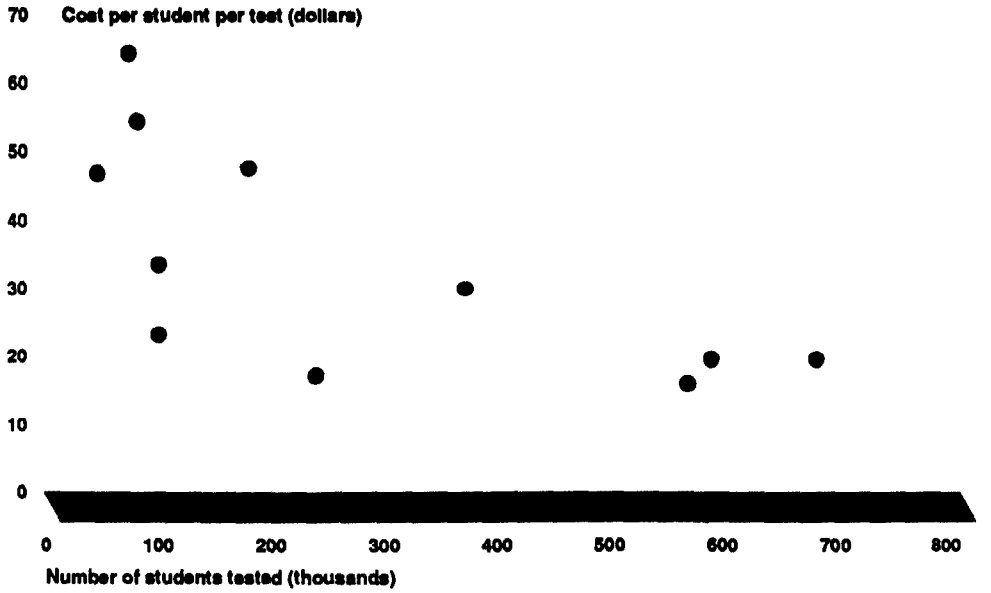
Figure 3.1: Per-Student Costs of Two Test Types in States Having Both



Economies in Testing

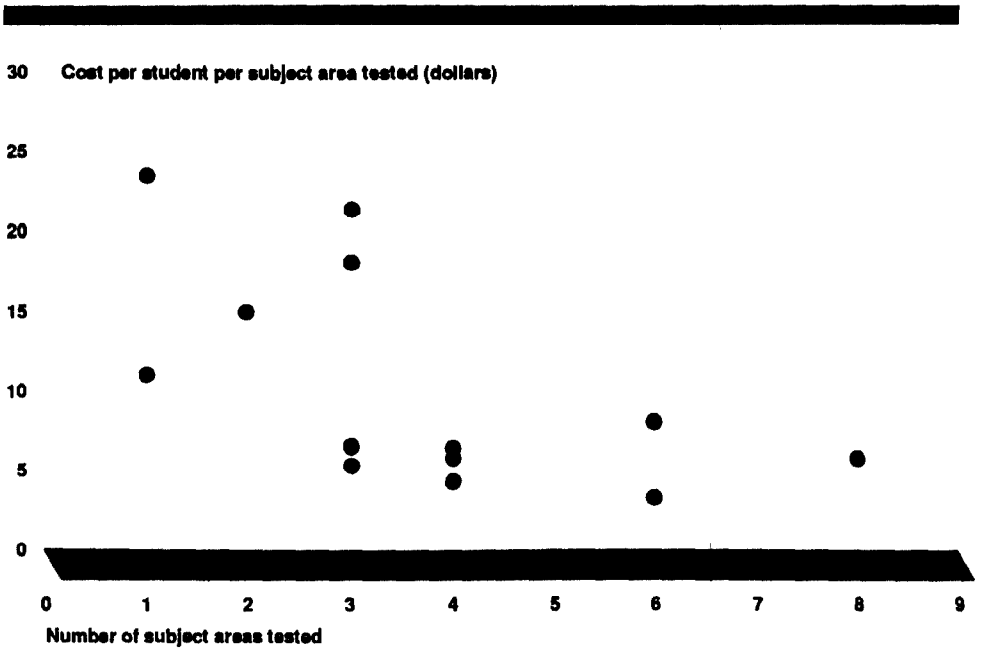
The cost of testing can be lowered over time, we found, through economies of scale and scope and as experience grows. This is especially important in considering the nationwide costs of performance testing, which has been very expensive in the pilot efforts so far. Our survey data provide evidence of all three possibilities for future economies, though we did not try to use our observations to create corrected or adjusted estimates of the long-term costs of one or more systems of national tests because so many factors are uncertain. As shown in figure 3.2, in reviewing our data on state performance-based tests, we found economies of scale when tests were given to different-sized groups of students. The per-student cost of a test declined as more students were included in a test administration. We can expect the per-student cost to decline as fixed costs (such as test development and some costs of operation, such as scoring, distribution, and site preparation) are divided by a larger test population.

Figure 3.2: Economies of Scale in State Performance Testing



Second, as shown in figure 3.3, again using the state performance-based test data, we found economies of scope when the same test administration was employed for several purposes, such as to test the same student population in more than one subject area. Again, we can expect the per-subject-area cost of a test to decline as more subject areas are included in a test administration and the fixed costs are divided by this larger number.

Figure 3.3: Economies of Scope in State Performance Testing



Third, costs can decline with experience, as those involved find ways to accomplish tasks in simpler and less expensive ways. For example, the state and the Canadian province reporting the most years of experience in performance-based testing have average per-student performance-based test costs of less than $22, well below the overall average of $33.

Test Development Costs

The data we used to describe testing costs thus far include all the costs our state and local survey respondents could recall for the academic year, 1990-91. Thus, we learned of ongoing test development costs (which can themselves be considerable), but we did not get good information on start-up development costs, those encountered before a test's first administration. We interviewed officials at testing firms and in state agencies to learn more about these one-time-only costs. Some tests in use today, such as the commercially produced, national norm-referenced achievement tests, were developed decades ago. Their start-up costs, even if adjusted for inflation, would bear little similarity to today's costs. Technologies, procedures, and expertise have changed a great deal over time and so have test development costs.

The many state tests developed within the past decade offer some more recent information. These state efforts have ranged widely in complexity. In all cases, commercial firms have been employed to do some or most of the work. The least expensive efforts have involved commercial test publishers' adapting their existing achievement tests to state curricula or other state needs. The test publishers tap their existing test-item bank as a source of questions for a particular state test. Officials of testing firms told us their start-up development costs ranged from one to a few dollars per student.

A more expensive way to develop a test is to start from scratch, writing test questions that fit a state's curriculum or guidelines, then testing the draft on pilot groups of students and making further revisions in the text, procedures, and so on. All of the recent state-developed, full-battery, performance-based tests have been done this way. From officials in six states with these tests and two more states where they were being developed, we learned that costs for initial test development averaged $10 per student. These state testing officials also told us the amount of time needed to develop the 10 tests from scratch to the pilot-test stage averaged 14 months and to final form, 27 months. None of these states used an existing state curriculum to develop questions for the tests. All developed state curricula or the like simultaneously with the tests.[7]

Two very expensive and current state test development efforts cost around $30 per student. These prototypes—the first of which was piloted in 1991-92—exclusively use performance-based formats (no part of the test uses the less expensive multiple-choice format) and cover many subject areas, including vocational education, art, music, or foreign language.[8]

Summary

Based on returns from our national sample of school districts and state education agencies, we estimate that the average per-student test cost in the United States in 1990-91 was $15. Multiple-choice tests tended to cost less while performance-based tests tended to cost more, and testing costs varied widely from district to district. School personnel time devoted to

[7]Testing officials developed what they called "curricular frameworks," "valued outcomes," "skill specifications," or "objectives," less detailed than a true curriculum. Testing officials in the one state among the eight with an existing curriculum at first tried to use it in developing test items, but abandoned the effort when they found it too cumbersome.

[8]Performance assessment and new testing techniques are discussed in detail in U.S. Congress, Office of Technology Assessment, Testing in American Schools: Asking the Right Questions, OTA-SET-519 (Washington, D.C.: U.S. Government Printing Office, February 1992), ch. 7-8.

testing accounted for three-quarters of a test's cost, which was borne for the most part at the local district level, even for statewide tests. State and local roles in testing differed; states did more test development and training, and local districts did more test administration and student preparation.

Both our sample of state performance-based tests and the national sample of all tests revealed economies of scale, scope, and learning. Some factors associated with higher testing costs included central city or rural school district location, low-income or ethnically-mixed populations, and state mandates to test.

Our surveys collected complete information about ongoing testing costs (including ongoing test development costs), but not for start-up costs incurred before the first test administration. A polling of state testing directors in states with the newer forms of statewide performance-based tests suggested an average start-up development cost of $10 per student and an average start-up development time of 3 years.

The Future Cost and Extent of Testing

In and of itself, current testing does not predict the future extent and cost of testing that would occur with the addition of a national examination system. That would depend on the type of national exams and on what states and school districts would do with their current tests. If they all were to keep all their current tests and add a national test, the extent and cost of testing would increase by an increment equal to the length and cost of the national test.[1] If schools were to replace an equal amount of current testing with a new national test, the extent and cost of testing would not change.

This chapter responds to the third evaluation question regarding the extent and cost of testing in a future with a national examination system and the overlap between current tests and those being proposed. First, we estimate the cost of a national examination system. Then we examine how local school districts reacted when their states mandated the use of new statewide tests, specifically, did they simply add the new tests or did they replace a then current test with the new state test? How they reacted provides a clue as to how state and local officials may react to a national test mandate. Finally, we discuss the impact of the type of national test, which may determine how many school districts replace current tests with a national test, how much additional testing time and expenditure will be required, and to what degree different school districts benefit from (or pay for) the change.

Cost of a National Examination System

To estimate the cost of a national examination system, many assumptions must be made about the type and extent of that system, especially the kinds of tests involved. Most recent discussions have proposed testing students at three grade levels. That would involve approximately 10 million students.[2] Given the range in test cost per student of $16 for multiple-choice tests to $33 for performance-based tests, as estimated in chapter 3, a national examination system could cost between $160 million and $330 million per year. As most recent discussions have proposed that a national system be made up of performance-based tests, however, $330 million may be the more relevant estimate. At a maximum, if the performance-based tests used in a national system were to cost as much as

[1]That is, the total cost would increase. The per-unit, or per-student-test, cost could possibly decline owing to economies of scale.

[2]Many national test cost estimates have used this figure in calculations, and it is plausible, as it represents one quarter (3 of 12 grades) of the nation's 40 million total enrollment in pre-college public education.

the most expensive state-developed performance-based test, the cost would rise to $640 million for a national system.[3]

Again, these estimates involve all testing costs, including the cost of personnel time spent in test-related activity. In other words, our figures include what it would cost at local and state levels to prepare, administer, and score a test given nationwide to all students in three grades, including paying for the time of all education personnel involved.[4] The estimates vary dramatically from both the high estimates offered by some national test opponents to the low estimates offered by some proponents.

High Estimates

The high estimates (at over $3 billion) were based on the per-student price for tests in five subject areas of the Advanced Placement examinations now administered by Educational Testing Service (ETS) for the College Board. Though the elaborate centralized marking of these long written exams is undeniably expensive, using these existing tests as a benchmark produces a high estimate for several reasons. The figure of $65 per subject-area exam is the price currently charged each student taking an exam, not the cost. Thus, some prior development costs may not be reflected (which understates the cost), and some current expenses paid from the fee may be for unrelated activities, such as fee reductions for low-income students, teacher training, or other activities of the College Board or ETS (which overstates the cost). Further, the five exams are separate, with five different administrations, each taking 3 hours. And the ETS staff told us that the $65 the student must pay is, in fact, an average price. Some Advanced Placement tests cost ETS more than that (art and foreign language exams) and some cost less (core subject area exams).

Low Estimates

Lower estimates of the cost of a national test have been made using the analogy of the Armed Services Vocational Aptitude Battery. This is a multiple-choice test composed of 13 subtests measuring abilities considered important for military service. It is administered by military personnel to all potential recruits and given free to school districts that wish to use it. Some inherent features of this test make it particularly weak

[3]The exceptionally high cost of that one test appears to result from the extra supervision believed to be necessary for administration of a high-stakes exam.

[4]We are not addressing the issue of who should bear the costs of national testing. All the costs, including the time of local district personnel, could be paid for nationally. Or, a test could simply be mandated nationally, with all costs left to local districts. In between these two extremes, the local districts might absorb the costs of test administration, while the test itself is developed and provided nationally.

as an analogy; for example, the inclusion of some topics such as electronics, the low degree of security, and the exclusive reliance on multiple-choice items. In addition, the commonly used cost data do not include the costs of staff administering the tests and analyzing and reporting the results.

Our Best Estimate

Generalizing from our survey data, the cost of all systemwide testing in the United States in 1990-91 totaled about $516 million. Given our best estimate of $330 million for a national examination system based on typical current performance-based tests, we believe a national system would cost almost two-thirds as much as the present cost of all systemwide testing.

By comparison, the annual federal contribution to local public schools has ranged between $7 billion and $10 billion in the past two decades. Total annual revenues to local public schools ranged between $110 billion and $130 billion over the same time period. Thus, annual total cost for a complete national examination system of the type we have been discussing would amount to less than 5 percent of present federal contributions to local public schools and to less than one-half of 1 percent of all government funds for local public schools.

We judged the current state performance-based tests to be a valid sample from which to estimate the cost of a national test for a couple of reasons. First, these tests resulted from consensus—from a political process with pressures and counterpressures similar to those one finds in the current debate over a national test. Different interest groups, testing experts, testing officials, and elected officials expressed concerns over test format, quality, cost, and length, and these are the tests they chose.

Second, these state testing programs have actually been implemented. So their extent and cost figures arise from actual practice, not as estimates. Because all but one of these programs are fairly recent, they may be more expensive now than they will be later, after testing officials have learned how to administer them more efficiently. Nonetheless, a national test could incorporate features that would make it more expensive. For example, a "high-end" national test could include only performance-based questions, which take more time to answer and to score, but it could still include enough questions to cover all subject area content thoroughly, and it could be a "high-stakes," and thus high-security, test. Extrapolating from the cost of a certain state test that if altered somewhat would resemble a

high-end test, we estimate that a high-end national test could cost over $1 billion.[5] No state or school district now administers a test with all the features of a high-end test, however.

Cost Economies

Economies in large-scale testing are relevant to any estimation of the cost of a national examination system, too. The decline in costs from having more experience in testing suggests that the cost of a national examination system should decline over time. The presence of economies of scope suggests that per-subject-area costs should decline as more subject areas are added to the same test. The presence of economies of scale suggests that per-student costs should decline as more students take the same test.

In the previous chapter, we noted that 11 state-developed performance-based exams, each covering four to eight subject areas, averaged $33 in per-student costs. Many advocate this type of exam as the main format for any national system of exams; thus, the $33 per student seems a reasonable estimate for the cost of national exams. If as has also been suggested, some of these current state performance-based tests were used by clusters of states, their per-student costs should decline as larger populations of students take each test and decline even more with experience over time.

Could economies of scale be pushed to an extreme, such that a single national test could be the least expensive—and most efficient—approach to broadening current testing? Our data do not suggest that. We found that the economies in performance-based testing seem to be exhausted at a much lower scale than that of the entire nation—at about the scale of a large state. Thus, grouping small states together in a cluster in which all use a common exam would achieve most or all of the possible economies of scale.

Start-Up Development Costs

Projecting start-up test development costs to the national level involves, once again, multiplying the per-student costs by the assumed 10 million

[5]This state test is completely performance-based in format, has high security because of that format, covers six subject areas, and employs local teachers and administrators in test development and scoring. There are currently three different forms of the test, and any one student gets only one of the forms. One form takes about 10 hours. Collectively, the three forms cover the entire curriculum; singly, each form covers only one-third of it. The test is now a "low-stakes" test. If the test were to be "high-stakes," in fairness, each student should be tested with the same test and over the entire curriculum. Such a test would take 30 hours. The current state test, in its first year of use, cost $48 per student. Tripling the test's length would not quite triple its cost, because ongoing development costs would not change. Moreover, a national administration of the test would benefit from economies of scale and, over time, economies of experience. Adjusting the cost estimate for economies of scale, a national high-end test could cost over $1 billion.

students. For the three methods of test development mentioned in chapter 3—for a multiple-choice test, for an average performance-based test, and for the most expensive type of performance-based test—national costs would amount to about $20 million, $100 million, and over $300 million. As most recent discussions of a national exam system have proposed the type of exam represented by the middle figure—a performance-based test like those currently in use in some states—$100 million probably ranks as the best estimate. Again, this represents a one-time-only cost that could be used to develop a new test or perhaps to pay the states that have already developed appropriate tests to share their knowledge. Table 4.1 summarizes these projections.

Table 4.1: Projected Costs of National Testing Options

	Type of test	
Testing cost	**Multiple-choice**	**Performance-based**
Per-student		
Start-up development	$2	$10
Annual administration	16	33
National (millions)		
Start-up development	20	100
Annual administration[a]	160	330

[a]Includes ongoing, recurring development costs.

The Effect of Adding a New Test

We examined districts' responses to past testing mandates as a way to estimate future responses to any required national test. Twenty years ago, very few states mandated statewide testing, but by 1990-91, only a handful of states remained that did not require their local districts to administer a statewide test. Twenty-five states required their districts to administer a commercially developed norm-referenced test. Thirty-three states have developed their own tests, either adapting an available commercial test to their needs (typically producing a criterion-referenced multiple-choice test) or developing their own test from scratch (usually producing a criterion-referenced performance-based test).

At the time their states mandated new tests, officials in local districts that were already testing faced a choice. They could replace an existing test with the state-mandated test and thus hold to the same number of tests, or they could add the state-mandated test to their testing program. Evidence from our surveys indicates that, in making their choice of whether or not to drop a test, local school officials considered the state-mandated test's

similarity in purpose and content to their existing test. As shown in table 4.2, we found that when the new state-mandated test was very similar, district officials tended to drop their existing local test. They were much more likely to keep their own test and add the new one when the state test differed in purpose or content.

Table 4.2: School District Responses to State Testing Mandates

State and local tests' purpose and content	Districts substituting state test
Exactly the same or very similar	82%
Somewhat or moderately similar	69
Not at all similar or very little	41

Still, 41 percent of districts dropped their own test even when it was different from the state's. The most common reasons cited by our survey respondents were that the new state test made their overall testing program too large or the new test was of higher quality than the old.

In conversations with officials from commercial testing firms and from our systematic sample of 50 school districts affected by state testing mandates, we learned that school districts try to spread the testing burden evenly across grade levels. When school districts add state-mandated tests at certain grade levels, they often move other tests to other grade levels. Some school districts augment the information gained from the state-mandated commercial tests by administering the same tests at other grade levels, at their own expense.[6]

This predilection of school districts to even out the testing burden across grade levels is corroborated in our national sample. As was mentioned in chapter 2, statewide tests tend to be concentrated at grades 3, 4, 6, 8, and 11. But districtwide tests, which include both statewide and exclusively local tests, are spread fairly evenly across the grades. So exclusively local tests are concentrated in the grade levels in which state tests are not administered.

[6]In our systematic sample of 50 school districts adopting state-mandated tests, 25 of them simply dropped an old test and 4 kept an old test in some of the same grade levels as the new state test. However, 15 districts kept an old test but moved it to grade levels not covered by the new state test. Six districts paid to supplement the state-mandated test in nonmandated grade levels.

Three Alternatives for a National Examination System

We projected the costs and other effects of three hypothetical alternative national testing plans, drawn from current debates. For example, the congressionally mandated National Council on Education Standards and Testing (NCEST) reviewed three main options in its work in 1991-92.[7] The several possible structures for a national examination system that NCEST considered can be summarized by three general types: a single national multiple-choice test; a single national performance-based test; and several clusters of states, each administering a different performance-based test. NCEST finally recommended the latter structure, which could employ several national performance-based tests, leaving the states free to choose among them. A "cluster" would be formed when several states decided on a particular test or developed one themselves. Conceivably, some or all of the current state-developed performance-based tests could be incorporated in a cluster system.[8]

Possible Responses to Each Plan

Knowing how districts responded in the past to mandated tests similar or dissimilar to those already in use, we assessed what could happen under each of the three alternative national test scenarios just described. From knowing past behavior in dropping tests or not and how many districts have tests similar to those proposed, we derived estimates of how many districts would replace current tests and how many would increase their testing programs by not replacing any current tests. From these data we derived further estimates of overall increased cost and testing time. The details of our procedure are shown in appendix II; the results are shown in table 4.3.

[7]By Public Law 102-62 (signed June 27, 1991), the Congress created NCEST to report by January 1992 on the desirability and feasibility of national standards and tests and on planning an appropriate system of tests. At the time, several independent groups had already proposed structures for a national examination system. In its first several months, NCEST studied those proposed structures and others generated by its own members.

[8]NCEST, Raising Standards for American Education (Washington, D.C.: January 24, 1992).

Table 4.3: School District Responses to Three National Test Alternatives

	Test alternative		
Response	**Single multiple-choice**	**Single performance-based**	**Clusters of performance-based[a]**
Add new test; drop old test	74%	52%	30%
Add new test; keep old test	26%	48%	43%
Marginal effect			
Additional annual cost (millions)	$42	$209	$193
Additional testing time (minutes per year per student)	15	30	25

[a]Our calculations assume that under a cluster system, 27 percent of school districts that now or soon will administer state performance-based tests will incorporate them into the cluster system.

We found that a single national multiple-choice test, which would most overlap with current testing, would add the least new time and money cost, as 74 percent of districts would drop some existing test. We estimated earlier in the chapter that the total absolute cost of a national multiple-choice test would be $160 million. Considering the 26 percent of districts that would maintain their existing test while adding the new national test, we estimate that only $42 million would be new costs. The remaining $118 million is already being spent, but on tests that would be replaced. We estimated only a small change in overall testing time per student per year, about 15 minutes more.

Because performance-based tests are much less commonly used, they would bring something new to many more districts, and as we have shown, districts facing such a choice are much more likely to add a mandated test that is different without replacing an existing test. Thus, as table 4.3 shows, we estimated that from 43 percent to 48 percent of districts would add a national or cluster performance test without replacing any current test, thus yielding a higher level of new costs—from $193 million to $209 million—and between 25 minutes and 30 minutes more testing time per year for the average student.

The single national multiple-choice test emerges as the least expensive alternative for two reasons. First, multiple-choice tests are inherently less expensive than performance-based tests to administer and to process. Second, they impose fewer new costs because they duplicate current testing the most.

Incidental Costs and Benefits of Adding a National Test

Replacement Disruption

The addition of a national test will affect more than the extent and cost of testing in the United States. It could disrupt present systems and testing programs. For example, school districts that drop currently used multiple-choice tests in favor of a new national multiple-choice test would give up some test familiarity, trend data, and perhaps a curricular alignment with the test. And if enough school districts abandon commercial and state-developed multiple-choice tests, some test developers could lose their jobs.

Similarly, school districts that drop currently used state performance tests in favor of a single national performance-based test would give up some test familiarity, trend data, and perhaps a curricular alignment with the test. Moreover, many state testing officials who now develop and administer state performance-based tests might find their jobs obsolete.

Presumably, with a cluster system, states would be able to join a cluster with a test that closely matches their curriculum, if they have one. So little would be lost in curricular alignment. Moreover, if states that currently administer their own performance-based tests were allowed to keep them by starting a cluster, no state testing officials would be displaced.

Windfall Benefits and Added Costs

Adding a test to a school district's testing program can be viewed as a benefit or as a burden—a benefit if the test is wanted, a burden if it is not. When a school district receives a desirable test free, as it might by participating in a national examination system, it receives a windfall benefit. It gets a test it wants without purchase or development costs. If the school district's administrative time costs were subsidized as well, the test would be an even greater benefit.

When a higher level of government mandates that a local school district administer an unwanted test, the test would create added costs to the school district unless the personnel time in administering the test was completely subsidized and did not detract from regular instruction (in which case, the new test would have a neutral effect). In most cases, a new test adds both benefits and costs to a testing program.

Efficiency Benefits

Efficiency benefits can result from the achievement of one or more economies in testing, such as those mentioned earlier—scale, scope, or learning. One might think that conversion to single national tests, whether multiple-choice or performance-based, would produce economies of scale. But scale economies in testing seem to be exhausted at a scale smaller than the whole country. Our analysis earlier in this chapter showed that scale economies in state performance-based testing seem to be exhausted at about the level of a large state. The fact that three profitable test publishers coexist selling several different nationally normed multiple-choice tests suggests scale economies might be exhausted for multiple-choice tests as well. (Otherwise, one of the three companies could undercut the other two on price, enlarge its market share while lowering its costs, and drive its competitors from the market.)

Performance-based tests are now being developed and administered by some relatively small states, and the larger scales of either a single national performance-based test or a cluster system should engender some scale economies.

A cluster system could produce some learning benefits. Several tests in separate clusters would, essentially, compete with each other for the allegiance of states, who would be free to select the clusters of their choice. Competition among the several clusters should provide incentive for them to learn how to lower costs and improve quality in order to retain the states within the cluster and to attract others. Learning effects can be particularly important with new or relatively undeveloped technologies. The more prevalent the opportunities to experiment with new methods, the faster the technology can develop.

Test Matching Costs

In and of itself, a national system of state performance-based tests would not provide comparability of test scores across clusters. For example, the median student test score in one cluster of states might not represent the same level of academic achievement as the median student test score in another cluster of states. The two clusters might have very different tests or tests that differ in their level of difficulty. Arranging the tests to produce comparable scores will require some effort and coordination. If such test-matching is to be done, it should be considered as a cost, one unique to the cluster design.

Table 4.4 summarizes the incidental costs and benefits of the proposed national tests.

Table 4.4: Incidental Costs and Benefits of Proposed Tests

	Proposed testing		
Current testing	**Single national multiple-choice test**	**Single national performance test**	**Clusters of performance tests**
States with no state test	Add a test; windfall benefit[a] and added cost[b]	Add a test; windfall benefit and added cost	Add a test; windfall benefit, added cost, and test-matching cost[c]
States with multiple-choice tests	Do not add a test	Add a test; windfall benefit and added cost	Add a test; windfall benefit, added cost, and test-matching cost
	Replace a test; replacement disruption[d]	Replace a test; replacement disruption	Replace a test; replacement disruption and test-matching cost
States with performance tests	Add a test; windfall benefit and added cost	Do not add a test	Do not add or replace a test
	Replace a test; replacement disruption	Replace a test; replacement disruption	Join a cluster; efficiency benefits[e] and test-matching cost

[a]Windfall benefit: State gets a new test to use free of some of its costs.

[b]Added cost: New test would be administered along with old test (or for the first time).

[c]Test-matching cost: Effort required to make scores comparable across clusters.

[d]Replacement disruption: In replacing an old test, a district may give up familiarity, trend data, or curricular alignment; commercial test publishers may lose customers; and employees managing state tests may no longer be needed.

[e]Efficiency benefits: Clustering helps small states as a group to reach efficient scale in testing.

Summary

Looking at a sample of states with performance-based tests much like those proposed for a national examination system allowed us to estimate the cost of a national system. Our best estimate is $330 million, and different alternatives could cost from $160 million to $640 million, far lower than the estimates of some national test opponents, and far higher than those of some proponents. Economies of scale, scope, and learning imply that the cost of any national system of exams should decline over time.

In general, when a school district in our sample adopted a mandated state test, it was more likely to abandon an existing test if the two were similar in purpose or content, and more likely to retain an existing test if the two were different. This suggests that the purpose or content of a voluntary national test might determine the degree of overlap with existing testing

programs. If the national test is different from existing tests, a district (or state) may just add the national test. If the national test is similar to an existing test, a district (or state) may jettison the existing test or not adopt the national test, effectively not enlarging its testing program.

Each of three alternative plans for a national examination system would likely have different effects on the extent and cost of testing in the United States. A single national multiple-choice test would likely replace tests now in use in three-quarters of U.S. school districts (90 percent of which would be other multiple-choice tests) and would add $42 million overall and 15 minutes per student to the current cost of testing ($516 million and 3.4 hours per student per year). A single national performance-based test would likely replace tests now or soon to be in use in just over half of U.S. school districts (42 percent of which would be other performance-based tests) and would add $209 million and 30 minutes of testing per student. A "cluster" system of performance-based tests would likely replace tests now in use in 30 percent of U.S. school districts and would add $193 million and 25 minutes of testing per student.

Testing Officials' Views on the Benefits and Costs of Present and Future Testing

The views of local and state school administrators on school testing can be important for several reasons. First, the administrators implement the present school testing programs and will determine much of the character of any new ones. Second, they are in a position to make informed judgments about the value of their current tests. Third, for some of the information we were asked to obtain, such as the benefits of testing, there is virtually no other practical way to get it.

This chapter presents a summary of the views of state and local testing officials on the benefits and costs of their testing programs, their perspectives on future trends in testing, and their reactions to the concept of a national exam or national examination system.

Benefits and Costs of Current Tests

Two survey questions addressed the net benefits (total benefits minus total costs) of local and state testing programs. Respondents from both groups strongly believed that the net benefits of their present testing programs were positive. Seventy-five percent of state respondents felt that way (compared to 5 percent who felt the opposite) and 43 percent of local respondents felt that way (compared to 18 percent who felt the opposite).[1] Those local district respondents who were testing directors were almost twice as likely as local district superintendents to see their testing program's net benefits as positive, though even superintendents leaned strongly in that direction. All our state respondents were either testing directors or administrators in the testing programs.

State respondents believed strongly that net benefits would increase if their testing programs were somewhat larger—52 percent indicated so (compared to 5 percent indicating the opposite).[2] But a larger state testing program necessarily means larger district programs, unless district administrators jettison an existing test when their state mandates a new one. At the local level, slightly more respondents (28 percent versus 22 percent) thought net benefits would decrease than thought they would increase with a somewhat larger district testing program, but 40 percent

[1]An additional 16 percent of state respondents and 34 percent of local respondents thought that the benefits and costs of their testing programs were about equal. Five percent of each group replied with "don't know or no opinion." The exact wording of the question was, "Do you believe that the benefits of your state's/district's present testing program are greater than the costs, that the costs are greater than the benefits, or do you believe that they are about equal?"

[2]Twenty-three percent of the state respondents felt that the net benefits would remain about the same if their testing programs were somewhat larger, while another 21 percent replied "don't know or no opinion." The exact wording of the question was, "Do you believe that the net benefits (total benefits minus total costs) to your state/district would increase if your testing program were somewhat larger, would decrease, or would remain about the same as now?"

thought net benefits would remain the same. Thus, 62 percent of local respondents thought net benefits would increase or remain the same with an additional test.

When asked in open-ended questions to list their testing programs' chief benefits, the local respondents overwhelmingly mentioned such benefits as diagnosis and evaluation information for students, parents, schools, programs, or districts. By contrast, other potential benefits, accounting for less than 15 percent of the responses, concerned positive classroom outcomes (improved student performance, curriculum alignment with standards, and so on), positive products of the assessment process (clear standards, better public understanding, teacher edification, and so on), or accountability at any level.[3] Clearly, local school officials viewed tests as helpful diagnostic instruments, though not clearly linked to practice or results, even while others perceived different purposes for the same tests.

State respondents were more likely than local respondents to mention other types of benefits, such as accountability (33 percent) or maintenance of common, clear standards (11 percent). Otherwise, they mentioned most often the diagnostic benefits. When asked to list the chief costs of their testing programs, the local respondents referred usually to direct costs, such as the test purchase, administration time, or scoring fees, or to opportunity costs, chiefly the loss of teaching and learning time. Though the question directed respondents to think of costs in a broad sense, few mentioned such problems as teaching to the test, misuse of test results, or stress. Much the same was true among the state respondents, though with them, test development was also often mentioned as a cost.

The Future of Testing

Majorities among both the local and state respondents saw more tests in their future, whether they liked it or not. Fifty-nine percent of the local respondents anticipated more state-mandated tests in particular, despite the fact that all but a few states already use at least one state-mandated test. Forty-six percent of the local respondents and 61 percent of the state respondents believed that the proportion of their education budgets devoted to testing would grow in the near future (compared to 6 and 2 percent, respectively, who believed the opposite). Plus, very large

[3]"Accountability" was defined in the questionnaire to mean that "assessment [is] used to determine promotion, retention, or graduation" at the student level; that "results are used to help determine principal's retention, promotion, or bonus, or cash awards to, honors for, status of, or budget of the school" at the school level; that "information is made public and voters or school board can instigate systemwide change" at the district level; and that "information is made public and voters or legislature can instigate systemwide change" at the state level.

majorities from both groups believed their testing programs were secure; that is, no more susceptible to budget cutbacks, or even less so, than other educational programs.

A majority of local respondents believed that the proportion of tests that are commercially developed will not change in the near future. A clear majority of state respondents, however, believed that they will rely less on commercially developed tests (only one state respondent expected to rely on them more). Both state and local respondents identified a trend toward more use of criterion-referenced and away from norm-referenced tests and a trend toward more use of performance-based and away from multiple-choice formats. Very large majorities among the state respondents confirmed the two trends—70 percent predicted more criterion-referenced tests (compared to 2 percent predicting more norm-referenced tests), and 87 percent predicted more performance-based tests (compared to no state respondents predicting more multiple-choice tests).[4]

When asked in open-ended questions to identify the most positive contemporary trends in student assessment, state testing directors mentioned most often: more performance-based and "authentic" tests (52 percent); improved testing procedures in general, such as criterion-referencing, less cultural bias in test items, and testing "higher order skills" (38 percent); and testing as part of integrated educational programs (11 percent). When asked to identify the most negative contemporary trends, they mentioned most often: misuse of test results to compare districts or states that are not alike as if they were or to make unwarranted inferences about students (47 percent); use of unproven methods (for example, performance-based tests—25 percent); and too much testing or too much emphasis on testing (21 percent). Table 5.1 summarizes these responses.

[4]In addition, 17 percent of state respondents predicted about the same proportion of criterion-referenced to norm-referenced tests, while 11 percent replied "no opinion or not applicable." Eleven percent of state respondents predicted about the same proportion of multiple-choice to performance-based tests, while 2 percent replied "no opinion or not applicable."

Table 5.1: Positive and Negative Trends in Testing

Trends	Respondents[a]
Positive	
More performance-based tests	52%
Improved testing procedures (less bias, higher order skills tested)	38
Testing as part of an integrated educational system	11
Negative	
Misuse of test results	47
Use of unproven methods (including performance-based tests)	25
Too much testing or emphasis on testing	21

[a]State testing directors responding to our survey. Up to three responses were counted from each respondent, so the percentages of all responses do not total 100.

Reaction to a National Examination

Our questionnaires were developed in 1991 when many proposals centered on a single national test, so we asked respondents for their reaction to "a voluntary national achievement test." Thus, their responses reflect their reaction to the idea of a single test. Were the questionnaire written today, that phrase might have been altered to read "national examination system," to better reflect the widely discussed recommendation of NCEST against a single test and in favor of a system incorporating several different tests. Some others who were opposed to a single test might favor such a "cluster" system of exams. We did, however, attach a series of open-ended questions to our survey that referred to a potential "national examination system."

The survey, then, asked respondents which factors, among 12 posed, would be most important to them if it were their responsibility to choose to adopt a voluntary national test. Among the 12, 3 factors were considered the most important by both groups: the quality of the national test, the cost to the state or district of administering the test, and the usefulness of the test results to state or local internal evaluations. Judging from their responses to other survey questions, we believe that our respondents would consider a test to be of higher quality to the degree that it covers what their teachers teach and measures diverse skills, by including some performance-based response formats as well as content-based, multiple-choice formats.

Other factors considered important, but less so than these three just mentioned, were those involving the fit between a national test and existing district or state tests. Kinds of fit that respondents said were

important included, in order of decreasing importance, similarity in content, purpose, grade level, test type, or time of year when given. A national test would be less accepted, that is, to the degree that it differed from current practice on these dimensions.

State respondents noted that if it were their responsibility to choose to adopt a voluntary national test, they would find it extremely important if the national test proposal were accompanied by pressure to adopt or not adopt from forces outside, such as the governor, the state legislature, or public opinion. Local respondents judged these considerations somewhat less important. When asked which factors they would consider most important in deciding to drop an existing test in favor of a national test, the relative ranking of factors mirrored that for the previous question on the simple adoption of the national test.

Separate, open-ended questions that asked for the perceived advantages and disadvantages of a "national examination system" revealed a good deal of opposition to the idea (see table 5.2). Forty percent of the local district respondents and 29 percent of the state respondents offered that there were no advantages or that they could not think of any. One positive advantage mentioned by a sizable number of respondents (over half of the state respondents and 32 percent of the local respondents) concerns the common metric and basis for comparison of performance that a national testing system could provide.

Table 5.2: Advantages and Disadvantages of a National Examination System

	Officials responding[a]	
Response	**State**	**Local district**
Advantages		
No advantages or cannot think of any	29%	40%
Common bases for comparison, clear standards	53	32
Disadvantages		
Misuse of test results	41	26
Push for national curriculum or a decrease in local control	25	14
Mismatch of test to local curriculum	20	4
Teaching to the test or a narrowing of curriculum	16	17
Use of restrictive or narrow testing formats	14	7

[a]Up to two responses were counted from each respondent, so the percentages of all responses do not total 100.

The potential linkage of such expanded testing to a national curriculum, the clear decrease in local control, and a lack of match of the tests to local curricula were mentioned often as disadvantages of a national exam system. Other disadvantages often mentioned concerned misuses of tests in general—not necessarily just a national test—such as the inappropriate comparison of unlike districts or states, inaccurate reporting of test results, teaching to the test, narrowing the curriculum, and use of restrictive or narrow testing formats.

A Trade-Off Between Test Quality and Cost

Although our survey respondents did not seem opposed to more testing, they were particular about the kinds of tests they favored. They indicated a preference for performance-based tests with the content based on their state or local curricula and results that can serve local purposes, such as student, school, or curriculum diagnosis. This desire does not necessarily match present practice. As we discussed in chapter 2, most state respondents reported no required state curriculum in 1990-91. Overall, state respondents identified only 46 percent of their statewide tests as largely or perfectly aligned with their state curricula. Even so, curricula, whether state or local, whether specified or just ad hoc, do not vary so much that disparate school districts cannot still use the same textbooks, virtually all of which are sold nationally.

Because the kind of testing our respondents want is not exactly what they now have does not invalidate their wishes, however. Testing directors and local superintendents and administrators work within the constraints of budgets and state mandates and cannot always completely control the make-up of their testing programs. Besides, it seems logical that they would desire a positive addition to their present testing programs. A national multiple-choice test—the low-cost alternative—would be largely duplicative; a curriculum-based performance test would be, for most districts, something new.

Some would argue, moreover, that the present commercially developed multiple-choice tests already are national tests; they are designed and developed with information drawn from national samples of students in pilot tests, and then they are sold nationally. Some critics of these tests have argued that the test publishers do not update the material in these tests often enough, that their test security is often lax, or that they test only a narrow range of skills and do not challenge the students enough

even within that range. Still more criticisms have been leveled against these tests.[5]

To be fair, however, we note that a sizable minority (25 percent) of state testing directors saw the use of performance-based tests as a negative trend. We cannot be sure, but they may have said this because of the undeniable fact that multiple-choice tests do have some advantages other than cost over performance-based tests. First, because multiple-choice test questions can be answered quickly, many more of them can be answered within a given time period. Thus, multiple-choice tests can cover the content of a subject area far more quickly than can a performance-based test.

Second, because multiple-choice tests limit the domain of possible answers and only one is correct, scoring the exams can be done quickly and with near-perfect consistency. Machines score multiple-choice tests, and every test is scored the same way. Individuals score performance-based tests, and each scorer may have a different idea of which answer is correct, how it should be expressed, and what score certain answers should get.[6]

Regardless of the test format, some efforts to cut costs can threaten test quality. To save money, testing officials can update the content of tests less often, develop shorter tests or fewer forms of a test, use fewer teachers to score performance-based test items, or make no effort to tie the content of a test to the subject matter actually taught in the schools. These particular cost-saving efforts can threaten test quality by decreasing the degree to which individual test results genuinely and accurately represent a student's knowledge.

Summary

Our respondents generally told us that they believed the net benefits of their testing programs were positive and would increase or remain the same if more tests were added. Thus, our local district and state

[5]A wide variety of problems with tests were raised in testimony before the House Committee on Education and Labor in 3 days of hearings on the NCEST proposals. See House Committee on Education and Labor, Oversight Hearing on the Report of the National Council on Education Standards and Testing, serial no. 102-105 (Washington, D.C.: U.S. Government Printing Office, February 19, 1992).

[6]For example, serious problems of reliability surfaced in a recent evaluation of one state's portfolio assessment (that involved teacher ratings of selected examples of students' writing and math). In this case, standardized scoring conditions—a major preventive against unreliable ratings—may have been difficult to obtain as a large number of teachers took part and their training was modest. See Dan Koretz, et al., The Reliability of Scores From the 1992 Vermont Portfolio Assessment Program: Interim Report, Technical Report No. 355 (Los Angeles: UCLA Center for the Study of Evaluation, December 1992).

respondents seemed not to be opposed to more tests, though the local districts, where the tests are administered, may be closer to the saturation point than the states. Moreover, though both state and local respondents were open to more testing, they were particular about the type of tests: they worried about their quality, purpose, and locus of control over content and administration.

Very clearly, local districts told us they use tests—and believed they should use them—as diagnostic instruments, to assess and improve the performance of students, programs, schools, or districts, rather than as accountability measures. Our respondents have indicated a preference for well-designed tests that served local purposes, such as student, school, or curriculum diagnosis.

Finally, the survey revealed a large amount of opposition in fall 1991, particularly at the local level, to the concept of a national test or national examination system. Tempering that opposition was an acknowledgment by 53 percent of state officials and 32 percent of local officials that a national examination system could provide a common, clear basis for comparing academic performance across the United States.

Conclusions and Matters for Consideration

Conclusions

A National Examination System May Not Be So Costly

Our estimates for the cost of a national examination system are higher than those of some national test proponents, but lower than those of some opponents. Our best estimates for the most likely type of test show a national cost near $330 million annually, or about one-tenth the amount that some test opponents have suggested. Of this, we estimate that close to $200 million would be new costs, while the rest would be compensated for by replacing some current tests. Start-up test development could add a one-time cost of $100 million.

A national examination system would likely increase by up to 30 minutes the average amount of systemwide testing time per student, increasing the national average to 4 hours per student per year—an amount of time that still does not seem unduly burdensome, especially in view of the powerful potential information gains.

Some Opposition Exists to a National Test

Though our respondents did not seem opposed to more testing that met certain quality, utility, and reporting criteria, many expressed opposition to a national examination system. That is, they opposed a national examination system in the abstract without knowing its particular characteristics. This opposition should give pause to advocates of a national system who are counting on the cooperation and support of state and local education officials who will likely be the ones responsible for administering and preparing the students for the exams. If they remain opposed to the idea, either because no one has convinced them of its worth, because they see it as a useless or harmful imposition, or because they do not see themselves involved, success is less likely.

No One Plan Dominates the Others

No plan is a clear winner. Table 6.1 compares three alternative national testing plans on the three main criteria we examined (cost, overlap, testing officials' preferences) as well as on three others where they have obvious, well-established differences (familiarity of method, comparability of scores nationally, and alignment of test and curriculum). A single national multiple-choice test offers lower cost, strong comparability of scores and the most familiar methodology. A cluster system of performance-based tests overlaps less with present testing, but may better match the preferences of state and local testing officials and has more chance for

curricular alignment. The third option, a single national performance-based test, could provide stronger test score comparability than the cluster plan and potentially stronger influence toward national standards and curriculum. Obviously, preferences for certain criteria or for the alternatives will vary among teachers, other educators and officials, and the public.

Table 6.1: Evaluating the Three National Test Alternatives

	National examination system alternative		
Criterion	**Single multiple-choice**	**Single performance-based**	**Clusters of performance-based**
Cost	Not costly	More expensive	More expensive
Overlap with present testing	More	Less	Least
Testing officials' preferences	Least	More	Most
Methodology	Familiar	Less familiar	Least familiar
Comparability of test scores nationally	Strong	Good	Weak
Curricular alignment	Strong if curriculum is national	Strong if curriculum is national	Possible even with diverse curricula

Matters for Congressional Consideration

Involvement of State and Local Educators

If the Congress wishes to build support for a national examination system among teachers and state and local administrators, it should consider specific ways to encourage their involvement in the process of curriculum development, standard-setting, and test development, administration, and scoring. This would improve the likelihood of success of a national system as local teachers and administrators should be an integral part of any test administration.

Done this way, test development efforts can still try to benefit from the lower cost, adherence to common standards, curricular integration, and other potential advantages of large-scale assessment while trying also to overcome local fears and alienation. Teacher involvement in test development seems to strengthen teacher adherence to standards and

curricular integration and to relate testing to improvements in teaching and learning.

Involving state testing officials in the planning and execution of a national system could be advantageous for two reasons. First, officials in the many states with active and sophisticated testing programs have developed a great deal of expertise in large-scale testing and, thus, have much to teach anyone planning a national system. Their expertise in technical aspects of testing may be shared by many experts in universities and elsewhere. But their expertise in the implementation of large-scale testing programs involving different types of tests and involving several groups of stakeholders is shared by few others.

A second reason for involving state testing officials in the planning and execution of a national system is to benefit from their views on the most orderly transition to a national system. Many state testing programs are now well-established or soon will be. Several others are being planned. Some of these programs are large, sophisticated, and complicated, and a national system will, inevitably, affect them.

Ensuring the Validity and Reliability of Tests

If the Congress wishes to encourage the development of a well-accepted and widely used national examination system, it should consider means for ensuring the technical quality of the tests. Large-scale performance-based testing, in particular, is both popular and new—only one state performance-based test is more than 6 years old and only two are more than 3 years old. Its newness suggests that development of appropriate, valid, and reliable tests and of efficient methods for scoring them will require some trial, effort, and time. State performance-based test development periods ranged from just 1 year to 3 years. Creating a national system of any kind, however, will be an endeavor of unprecedented scope. Coordinating the efforts of several layers of government alone should challenge the best of planners.

Test quality will require an enduring commitment and sufficient resources to ensure that any tests in a national system are valid and reliable. Pressures to cut corners and degrade the quality of tests are inevitable. Money and time can be saved, at the expense of high quality, for example, by creating fewer forms of a test, forgoing pilot tests of test items, shortening the length of a test, or relaxing security. The need for quality controls is underscored by the views and preferences of the testing officials who responded to our survey. They prefer tests with high-quality

characteristics and they worry that a national examination will not embody those characteristics. And they worry that test results may be misrepresented and misused.

It is, however, beyond the scope of our study to suggest what means should be used to ensure quality in a national system of examinations. The National Council on Education Standards and Testing has proposed that a national technical panel be appointed for this purpose. There are other possible ways to ensure quality. In view of the sizable controversy over current testing, and the potential for incorrect decisions based on flawed test data, quality assurance in an expanded system is extremely important and should be explicitly and proactively considered in any national examination system implementation plan.

Sample Survey: Statistical Analysis

The representative national sample from which we derived our estimates on the cost, extent, and nature of testing in the United States consisted of 500 school districts. We received 368 completed questionnaires from this group, for a 74-percent response rate.

Differences in Survey Response Rates Among Respondent Groups

National estimates built up from sample survey data can be shaky if some groups surveyed did not send back many responses. We analyzed differences in survey response rates among groups based on all the characteristics for which we had information: metropolitan status of district (urban, suburban, or rural); district student population size, number of statewide tests in district's state, number of statewide criterion-referenced tests, and number of statewide performance-based tests.

Only the difference in response rates across districts with different student population sizes proved to be statistically significant. We used a statistical test called the chi-square, which signals the likelihood that a pattern of differences among groups would prove to be, upon further repetition, consistent. A lower chi-square statistic suggests a strong probability that the response rates among respondent groups are truly the same, and a higher chi-square statistic suggests a low probability that the rates are truly the same.

For district size, the chi-square was a relatively high 11.031, with a very small chance, a probability level of 0.004, that the response rates were actually the same among the respondent groups. As shown in table I.1, the other chi-squares were low, with corresponding high probabilities of truly similar response rates among the respondent groups.

Table I.1: Chi-Square Tests Comparing Survey Response Rates Among Respondent Groups

Respondent characteristic and groups	Chi-square	Probability
Metropolitan status (urban, suburban, rural)	2.35	0.308
District student population size (small, large, very large)	11.03	0.004
Number of statewide criterion-referenced tests (0, 1, or 2)	1.28	0.527
Number of statewide performance-based tests (0, 1, 2, or 3)	3.15	0.369
Number of statewide tests (0, 1, 2, or 3)	2.01	0.571

District student population was categorized in three sizes—small (lower than 3,500 students), large (between 3,500 and 35,000 students), or very large (more than 35,000 students). The response rates varied for the three groups—with a 69-percent rate from small districts, an 83-percent rate

from large districts, and a 79-percent rate from very large districts. The differences in response rates among the three sizes of districts do not imply a bias in the estimates toward the large and very large districts, because all the estimates were weighted. For example, one cell that represents 30 districts in the United States could be represented by 7 districts responding to our survey. Another cell that represents 300 districts in the United States could be represented by another 7 districts responding to our survey. In the first case, the responses from the 7 districts are given the weight of 30 districts in the national estimates, and in the second case, the responses of the 7 districts responding to our survey are given the weight of 300 districts in the national estimates.

The national estimates would not be biased in favor of large and very large districts, for the small districts are sufficiently represented in the national sample through the weighting. However, in this example the estimates for small districts would be less reliable (i.e., less accurate) because they would be based on a smaller percentage of the group.

The estimates we derived from the group of small districts should be reliable, however, because the response rate for the group was still rather high—69 percent. The group was large—214 school districts responded, out of 312 surveyed. And it was, almost certainly, the most homogenous (in terms of the extent and cost of testing) of the three district sizes. Small districts were well represented in the respondent group because there were so many in the original sample—312 of the original 500 were small districts. By contrast, only 42 of the original 500 were very large districts.

We cannot, of course, demonstrate that nonrespondents might not be systematically different on other (nonmeasured) factors. Again, however, the response rate was sufficiently high to mute such concerns.

Confidence Intervals on Key Estimators

Presented below in table I.2 are the 95-percent confidence intervals for the key variables in the report. The estimates are provided for the sample as a whole. Standard errors for all variables are available from our office upon request.

Table I.2: 95-Percent Confidence Intervals for Key Variables

Variable	Estimate	Lower bound	Upper bound
Average amount of hours spent taking test, per student per year	3.4	3.1	3.8
Average amount of hours spent in all test-related activity, per student per year	6.5	5.4	7.6
Average cost per test administration per student[a]	$14.51	$12.61	$16.41
Average purchase cost per test administration per student[a]	$4.33	$3.79	$4.87
Average personnel time cost per test administration per student[a]	$10.18	$8.56	$11.80
Total cost of testing nationwide in 1990-91[a]	$516 million	$448 million	$583 million
Total number of districtwide tests administered in 1990-91	35,600	32,700	38,500
Total number of individual test administrations in 1990-91	36 million	32 million	39 million

[a]Estimate includes state- and district-level costs. The confidence intervals, however, pertain only to the district-level costs. Our district-level estimates were derived from a sample of districts. Our state-level figures are totals from the universe of all the states and, thus, are not estimates at all.

Marginal Effect of Proposed Testing Over Current Testing

This appendix gives details of our analysis of how school districts would react to several national testing alternatives. It is based on responses to survey questions about how districts had responded in the past to state-mandated tests (see table 4.2) and on other information about districts' current tests.

Potential Response to a Single Multiple-Choice Test

If all school districts were to adopt a single national multiple-choice test, we estimate 74 percent of them would drop another test, thus not enlarging their testing programs. The remaining 26 percent would add the national test without dropping another test.

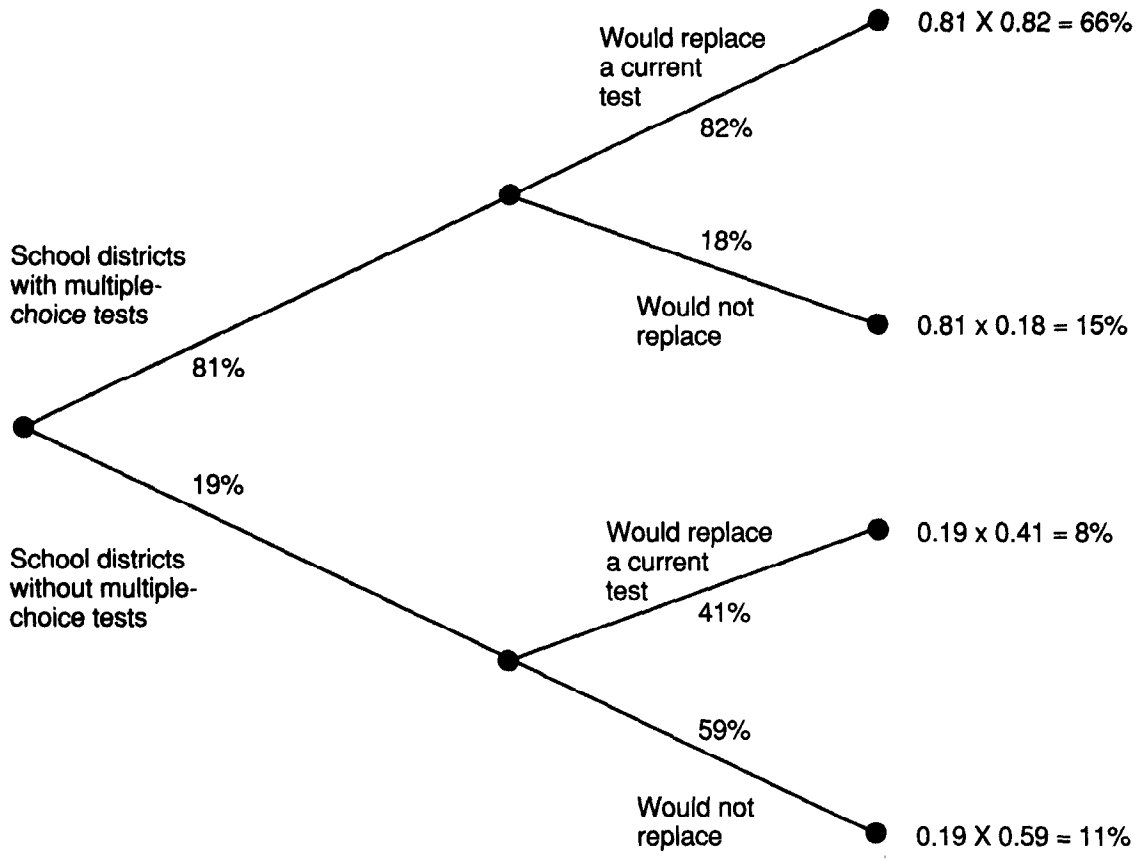
A single national multiple-choice test would clearly overlap in the 81 percent of school districts that now administer full-battery (multi-subject) multiple-choice achievement tests systemwide. Using the figure (from table 4.2) of 82 percent to estimate the proportion of districts with similar tests that would drop a current test, we conclude that 66 percent of all districts would do so (and 15 percent would not).

Similarly, using 41 percent (again, from table 4.2) as our estimate of the fraction of those districts that do not currently administer a multiple-choice achievement test but that would replace some current test with the national test, we see that another 8 percent of all districts would drop a current test (and 11 percent would not).[1]

This can be more easily visualized in a "tree" diagram of conditional probabilities as shown in figure II.1. The tree diagram shows that school districts, either with or without multiple-choice tests, might replace a current test, though the probabilities of that happening differ between the two groups. Adding the two different replacement probabilities (66 and 8 percent) together produces an overall replacement probability of 74 percent.

[1]That is, some current test other than a full-battery multiple-choice test. These are districts not currently administering full-battery multiple-choice tests, but administering other types of tests.

Figure II.1: Degree of Overlap With a Single National Multiple-Choice Test



School districts that would replace a current test = 66% + 8% = 74%
School districts that would not replace a current test = 15% + 11% = 26%

∴74% = replacement testing; 26% = new, additional testing

Though we estimated in chapter 4 that a single national multiple-choice test would cost around $160 million a year to administer, the analysis here shows that some of this cost would be new and some of it would be compensated for by school districts dropping old tests and their costs. Using our replacement probabilities, we calculate that a single national

multiple-choice test would add only $42 million a year in new costs. The new testing would also add about 15 minutes to the average of 3.4 hours per student in systemwide testing. The calculations are shown below.

First-Order Conditions

Cost: $16 per student per multiple-choice test

Number:

10 million students tested (3 grade levels)
40 million U.S. students total
26% of school districts adopting new tests

Time: 4 hours per test

Calculations

10 million x 0.26 = 2.6 million students
2.6 x 4 hours = 10.4 million new hours
10.4 ÷ 40 million = 0.26 hours (15 minutes)
0.26 hours x $16 per test x 10 million students = $42 million new costs

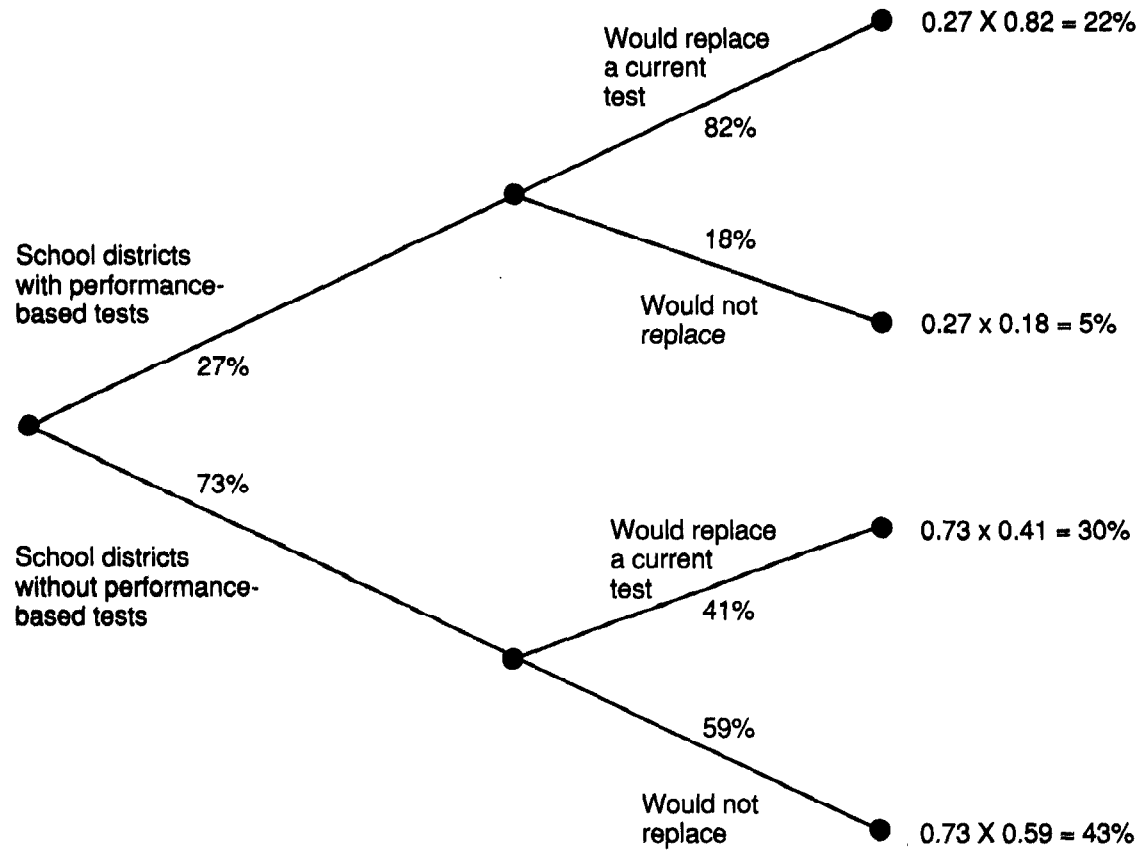
Of $160 million costs, $42 million new, $118 million replacement

Potential Response to a Single Performance-Based Test

Another tree diagram, figure II.2, illustrates the equivalent circumstances that would obtain if all school districts were to adopt a single national performance-based test.[2] Fifty-two percent of school districts would drop another test, thus not enlarging their testing programs. The remaining 48 percent would add the national test without dropping another test, thus adding to the extent and cost of testing.

[2]To make the numbers more relevant, we count among the school districts with performance-based tests all those in the nine states that plan to have statewide performance-based tests 3 years from now. Only seven states now administer statewide performance-based tests.

Figure II.2: Degree of Overlap With a Single National Performance-Based Test



School districts that would replace a current test = 22% + 30% = 52%
School districts that would not replace a current test = 5% + 43% = 48%

∴ 52% = replacement testing; 48% = new, additional testing

Though we estimated in chapter 4 that a single national performance-based test would cost around $330 million a year to administer, again the analysis here shows that some of this cost would be new and some of it would be compensated for by the districts dropping old tests. Using our replacement probabilities, we calculate that a single

national performance-based test would add about $209 million a year in new costs. The new testing would also add more than 30 minutes to the average of 3.4 hours per student in systemwide testing. The calculations are shown below.

First-Order Conditions

Cost $33 per student per performance-based test

Number:

10 million students tested (3 grade levels)
40 million U.S. students total
48% of school districts adopting new tests

Time: 4 hours per test

Calculations

10 million x 0.48 = 4.8 million students
4.8 x 4 hours = 19.2 million new hours
19.2 ÷ 40 million = 0.48 hours (30 minutes)
0.48 hours x $33 per test x 10 million students = $158 million new costs

Of $330 million costs, $158 million new, $172 million replacement

Because 30 percent of the $33 performance-based tests will replace $16 multiple-choice tests:

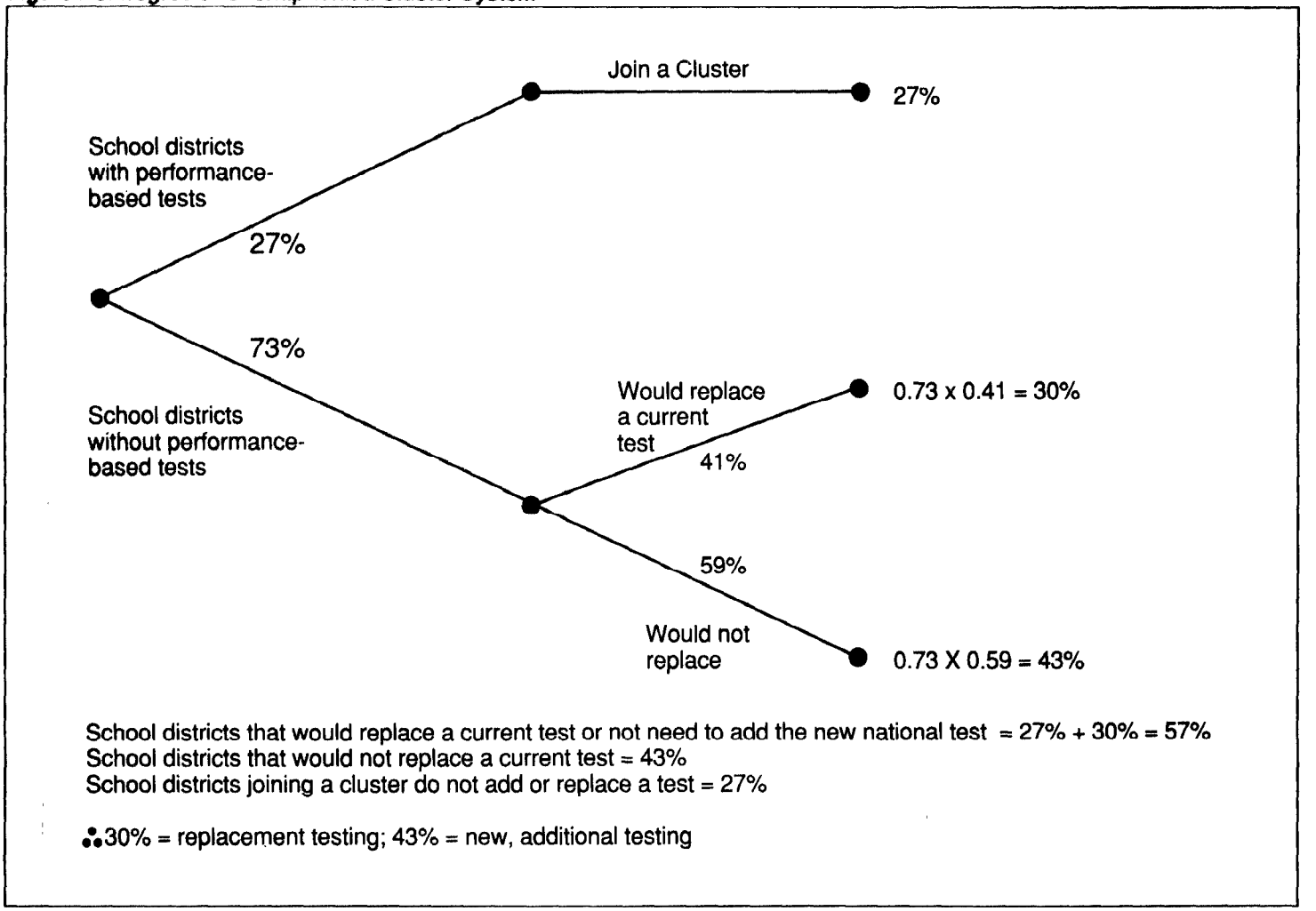
10 million x 0.30 = 3 million students
3 million x ($33 – $16) = 3 million x $17 = another $51 million new costs

$158 million + $51 million = $209 million

Potential Response to a Cluster System

The last tree diagram, figure II.3, illustrates the situation if all school districts not now administering state performance-based exams were to adopt a performance test from one of the national "clusters." Fifty-seven percent of the school districts would drop another test, thus not enlarging their testing programs. Forty-three percent would add a national test without dropping another test, thus adding to the extent and cost of testing.

Figure II.3: Degree of Overlap With a Cluster System



Using our replacement probabilities, we calculate that a cluster system of performance-based tests would add $193 million a year in new costs. The new testing would also add more than 25 minutes to the average of 3.4 hours per student in systemwide testing. The calculations are shown below.

First-Order Conditions

Cost: $33 per student per performance-based test

Number:

10 million students tested (3 grade levels)
40 million U.S. students total
43% of school districts adopting new tests

Time: 4 hours per test

Calculations

10 million x 0.43 = 4.3 million students
4.3 x 4 hours = 17.2 million new hours
17.2 ÷ 40 million = 0.43 hours (25 minutes)
0.43 hours x $33 per test x 10 million = $142 million new costs

Of $330 million costs, $142 million new, $99 million replacement

Because 30 percent of the $33 performance-based tests will replace $16 multiple-choice tests:

10 million x 0.30 = 3 million students
3 million x ($32 – $15) = 3 million x $17 = another $51 million new costs

$142 million + $51 million = $193 million

The Extent and Cost of Other Standardized Testing

To find the extent and cost of the most widespread tests, our surveys asked about systemwide testing done in U.S. schools. We defined systemwide tests as those administered to all students, almost all students, or a representative sample of all students in a school district in at least one grade level. Most standardized tests are given systemwide, but not all. Schools give some standardized tests only to certain groups of students. How much standardized testing did we miss by our choice of tests to study? We think not much. This appendix gives our estimates of the extent of three kinds of tests beyond those covered in our survey—tests given to meet evaluation requirements of the federal Chapter 1 program, state advanced achievement tests, and college entrance tests.

Chapter 1 Testing

Chapter 1 is the federal program providing supplementary services for economically disadvantaged students. Most school districts receive some Chapter 1 funds, which are targeted to schools that exceed a minimum percentage of economically disadvantaged students. Thus, Chapter 1 funds may support activities at some schools within a school district but not at others.

To identify educationally disadvantaged students for services and also to check the Chapter 1 program's effects, participating schools must test students both at the beginning and at the end of a reporting period. The test employed must be nationally normed. Because they have nationally normed tests readily available, commercial test publishers supply virtually all the tests used for Chapter 1 testing. Furthermore, because the publishers have formatted all the nationally normed tests with multiple-choice questions, Chapter 1 tests are always in multiple-choice format.

According to Department of Education officials, some school districts use the Chapter 1 testing requirement as an opportunity to test all their students at one or more grade levels. Thus, instead of purchasing just enough test booklets for the students in their Chapter 1 schools, district officials purchase enough test booklets for all the students in certain grade levels. That way, they obtain information on all their students and they fulfill their Chapter 1 evaluation requirement, paying a lower price than they would if they tried to meet both evaluation objectives separately. Department of Education officials believe that most school districts do their Chapter 1 testing this way, administering full-battery commercial tests to all students in certain grade levels in both Chapter 1 schools and non-Chapter 1 schools.

If the majority of school districts receiving Chapter 1 money do, indeed, test all students at the same time and with the same test as their Chapter 1 students, then most Chapter 1 testing is systemwide and is represented in the data the national sample of local school officials provided on our surveys. We have no way of precisely estimating how much Chapter 1 testing is systemwide and how much is not.

Even adding all Chapter 1 testing to our estimate of the extent of testing does not markedly increase our estimate, however. We calculated this extreme case, which assumes that no Chapter 1 tests were included in our surveys, so as not to underestimate the added testing burden caused by Chapter 1 evaluation. About 1.5 million students take Chapter 1 tests in reading and about 1 million students take Chapter 1 tests in math.[1] Department of Education officials told us the tests (given twice) take about 45 minutes per test administration. This amount of testing adds less than 6 minutes, or 0.1 hour, to our estimate of 3.4 hours of systemwide testing per student.

All systemwide testing and Chapter 1 testing comprise the group of all mandatory, school district-administered standardized academic tests. Our national sample of systemwide tests comprises 98 percent of the tests in this group.

State Advanced-Subject-Area Tests

In addition to statewide achievement tests, two of the larger states administer advanced-subject-area tests to some of their high school students. These are not systemwide tests because not all students in any one grade level take these tests, only those registered in certain advanced high school courses. The advanced-subject-area exams are administered to about 2.5 million students for about 3 hours each. From information provided in interviews with the two states' testing officials, we calculated the time involved for all students taking all the different subject tests, and found that in total those tests add 12 minutes, or 0.2 hour, to our 3.4-hours-per-student average for the extent of testing in the United States.

All systemwide testing, Chapter 1 testing, and these state advanced subject-area tests comprise the group of all school district-administered standardized academic tests. Our national sample of systemwide tests comprises 93 percent of the tests in this group.

[1]Beth Sinclair and Babette Gutman, A Summary of State Chapter 1 Participation and Achievement Information for 1989-90 (Washington, D.C.: Department of Education, 1992), p. 46.

College Entrance Examinations

The college entrance examinations of the American College Testing Program (the American College Test and Preliminary American College Test) and the Educational Testing Service (SAT, Preliminary Scholastic Aptitude Test, and the Advanced Placement exams) are not administered by school districts but by the testing firms, themselves. High school students are not required to take them; they take them only if they are considering applying to colleges and universities that require them for admission or advanced course credit. From figures supplied by the two firms on test times and number of students involved, we calculated that including all the nationally standardized college entrance exams adds 20 minutes, or 0.3 hours, to our national average of 3.4 hours per student for the extent of testing in the United States.

All systemwide testing, Chapter 1 testing, state advanced-subject-area tests, and college entrance exams comprise the group of all standardized academic tests. Our national sample of systemwide tests comprises 86 percent of the tests in this group.

Other Standardized Tests

The standardized tests for school-age students that remain are those given to special populations, such as psychological tests for special education students, IQ tests for gifted and talented students, or optional nonacademic tests, such as vocational-interest tests administered after school hours to students who elect to take them on their own time. We did not examine these. Compared to the national sample of systemwide tests, they are not many, and they are not like achievement tests, the kind of tests being considered for a national examination system.

Other Testing

Most tests, of course, are not standardized. Classroom teachers develop and administer most tests as a normal part of academic coursework. We know of no completed studies designed to accurately determine the extent of teacher classroom testing. And such a study was well beyond our resources to undertake.

Other Estimates of the Extent and Cost of Testing

This appendix summarizes other attempts to estimate the current extent and cost of testing in the United States. These studies have derived their estimates either from aggregate figures or from case studies.

OTA Estimates

The Office of Technology Assessment (OTA) did not attempt to estimate the current extent and cost of testing but did provide some pertinent information that we examined to see how consistent it was with our own data. The OTA report includes information from one large urban school district on all outlays for one school year on materials, services, and personnel related to standardized testing.[1] From data in the OTA report, we calculated that expenses on standardized testing amounted to less than one-half of 1 percent of the district's budget. That's a typical level of spending for large districts in our national sample.

OTA also reported the extent of testing in that district, finding the average student took 5 to 6 hours of standardized tests per year. This is slightly more than our national average of 3.4 hours per student per year. But we also found in our national sample that districts with some of the characteristics of the district OTA studied—central city location, a high level of poverty, and Northeastern location—had somewhat more testing hours.

The other test cost information in the report is derived from a report prepared for OTA by university researchers. They stated that performance-based tests in Great Britain and Ireland cost $107 per student, and OTA used this figure to represent potential costs of performance-based tests in the United States.[2] None of the state performance-based tests in our national sample cost that much (we found an average cost of $33 and a range from $16 to $64), though such a cost figure could be expected given certain conditions. The conditions surrounding the European tests were not specified in the researchers' report.

NCTPP Estimates

Between 1987 and 1990 the Ford Foundation sponsored the work of the National Commission on Testing and Public Policy (NCTPP), which centered chiefly on equity issues in the design and use of tests. Using some

[1]Testing in American Schools: Asking the Right Questions, OTA-SET-519 (Washington, D.C.: U.S. Government Printing Office, 1992), pp. 27-29.

[2]George F. Madaus and Thomas Kellaghan, Student Examination Systems in the European Community: Lessons for the United States. Contractor report submitted to OTA, June 1991.

estimates and some aggregate figures (for the reported sales revenue and volume of commercial tests, for example) the Commission's report estimated that "mandatory testing consumes some 20 million school days and the equivalent of $700 to $900 million in direct and indirect expenditures annually."[3] The report cites as its source a book that is as yet unpublished, so we could not determine how NCTPP made these estimates.

The Commission's figures are, nonetheless, close to ours. Using our figure of about 3.4 hours of testing for the average student, the approximately 40 million students in public elementary and secondary schools would spend in the aggregate 17 million 8-hour days on tests. Using a 6-hour school day in the calculation, we would estimate a somewhat higher total—23 million school days of testing per year. We also report in chapter 4 an overall estimate of $516 million in testing costs annually, which falls below the Commission's estimate, but we do not know exactly what they were counting as "indirect expenditures." The report used some very strong language to emphasize its contention that this amount of testing is "too much." The figure of less than one day per student per year, on average, seems not so alarming to us, but the conclusion is a matter of judgment.

The Commission report also estimated that students take 127 million tests per year, with individual students at some grade levels taking from 7 to 12 tests in a year. But in calculating these estimates, the Commission separated test batteries into their several subject-area components and counted each of them as a test. A typical commercial test of 4 to 5 hours in length might contain separate sections covering the basic subject areas of reading, grammar, math, science, social science, and writing. The test publisher and most others would still call it one test; the Commission described this as six tests. We estimate that U.S students take about 36 million tests per year.

Other Estimates

In our search of the literature, we found only two other empirical estimates of the extent or cost of testing that were based on reasonably complete calculations. A survey of school districts in 14 Northwestern states estimated that "the average student experiences 2 to 6 hours of

[3]From Gatekeeper to Gateway (Boston: 1990), p. X.

testing each year throughout elementary and secondary school."[4] That range includes our estimate from our national sample of 3.4 hours.

From a 1982 case study of one suburban school district, the Test Use Project at the University of California at Los Angeles calculated districtwide testing costs to be one-half of 1 percent of district expenditures, about the average that we found.[5]

[4]Beverly Anderson, "Test Use Today in Elementary Schools and Secondary Schools," in Alexandra K. Wigdor and Wendell R. Garner, eds., Ability Testing: Uses, Consequences, and Controversies (Washington, D.C.: National Academy Press, 1982), pp. 232-254.

[5]D. Dorr-Bremme and J. Caterall, Test Use Project: Costs of Testing (Los Angeles: UCLA Center for the Study of Evaluation, 1982).

Major Contributors to This Report

Program Evaluation and Methodology Division

Frederick V. Mulhauser, Assistant Director
Richard P. Phelps, Project Manager
Gail S. MacColl, Social Science Analyst
Christine Ing, Researcher
Cynthia S. Taylor, Researcher
Harry M. Conley, Sampling Consultant
Venkareddy Chennareddy, Referencer

Glossary

Accountability	Defined in our questionnaire to mean assessment that is "used to determine promotion, retention, or graduation" at the student level; whose "results are used to help determine principal's retention, promotion, or bonus, or cash awards to, honors for, status of, or budget of the school" at the school level. At the district level, "information is made public and voters or school board can instigate systemwide change," and at the state level, "information is made public and voters or legislature can instigate systemwide change."
Achievement Test	A test designed to measure a person's knowledge, understanding, or accomplishment in a certain subject area, or the degree to which a person possesses a certain skill. Achievement tests should be distinguished from aptitude tests, which attempt to estimate future performance.
Assessment	Generally refers to large-scale, systemwide measurement programs for pupil diagnosis, program evaluation, accountability, resource allocation, or teacher evaluation.
Criterion-Referenced Test	A test that allows its users to interpret scores in relationship to a functional performance level. Criterion-referenced measures provide information as to the degree of competence attained by a particular student, without reference to the performance of others.
High-Stakes Test	A test that is used to determine promotion, retention, or graduation. "High-stakes" tests and tests used for "student-level accountability" are considered synonymous.
Norm-Referenced Test	A test that shows a person's relative standing along a continuum of attainment in comparison to the performance of other people in a specified group, such as test-takers of a certain age or group.
Performance-Based Test	A test that measures ability by assessing open-ended responses or by asking a person to complete a task. Also known as alternative assessment, constructed response, or task performance, performance-based tests require the respondent to produce a response or demonstrate a skill or procedure. Examples include answering an open-ended question,

	conversing in a foreign language, solving a mathematics problem while showing all calculations, writing an essay on a given topic, or designing a science experiment.
Reliability	The reliability of a test refers to the degree to which test results are consistent across test administrations. Individual student scores are reliable if the same student gives the same answers to the same questions asked at different times. Test reliability can also be measured at the classroom, school, or district level. Tests tend to be reliable if their questions are clear and focused and unreliable if their questions are vague, contradictory, or confusing. Reliability can be measured rather precisely.
Representative Sample	A sample is a subgroup of a population. A sample is representative if it accurately reflects the character of the population in those aspects under study.
Standardized Test	A test is standardized if it is given in identical form and at the same time to students in more than one school, and all the results are marked in the same way. Tests scored by machine-reading of student marks in answer "bubbles" are not the only type of standardized test. Tests with open-ended essay questions and other kinds of performance-based tests can be standardized, too, if the conditions of administration and scoring are carefully controlled across schools.
Stratified Sample	In stratified sampling, a researcher selects randomly within each of separate homogenous subsets, or strata. The values derived from each of these subsamples are then weighted according to the proportion of the population represented by each subset.
Systematic Sample	In a systematic sample, the researcher randomly picks a number between zero and a number n/x, with n being the population size and x being the size of the systematic sample. Then, starting with that random number, the researcher picks every n/x item until x items are selected. In our study, we picked a systematic sample from our national sample, picking every n/x item in the order in which the questionnaires were returned in the mail.

Systemwide Test

We defined systemwide tests, for the purpose of this study, as any test that is administered to all students, to almost all students, or to a representative sample of all students within a jurisdiction for at least one grade level. Such a test can include several subject areas in a test battery. Tests that are optional for the student (as are college entrance tests) or that are only administered to unrepresentative subsets of the student population (as are tests for special education students) are not included.

Validity

The validity of a test refers to the degree to which it measures what it is designed to measure. There are several kinds of validity. Curricular validity, for example, would be strong if a test contained questions based on the content of the curriculum and weak if a test contained questions not based on the content of the curriculum. Predictive validity would be strong if an individual's test score accurately forecasted some other event, such as the likelihood of graduating or succeeding in a particular endeavor. Unlike reliability, validity is difficult to measure precisely.

Bibliography

Anderson, Beverly. "Test Use Today in Elementary Schools and Secondary Schools." In Alexandra K. Wigdor and Wendell R. Garner, eds., Ability Testing: Uses, Consequences, and Controversies. Washington, D.C.: National Academy Press, 1982.

Burry, James, et al. Testing in the Nation's Schools and Districts: How Much? What Kinds? To What Ends? At What Costs? Los Angeles: UCLA Center for the Study of Evaluation, 1982.

Caterall, James. The Cost of Instructional Information Systems: Results From Two Study Districts. Los Angeles: UCLA Center for the Study of Evaluation, 1983.

Caterall, James. "Fundamental Issues in the Costing of Testing Programs." In M.C. Alkin and L.C. Solmon, eds., The Costs of Evaluation. Beverly Hills, Calif.: Sage, 1983.

Coley, Richard D., and Margaret E. Goertz. Educational Standards in the 50 States: 1990. Princeton, N.J.: Educational Testing Service, August 1990.

Cronin, J.M. The Cost of National and State Educational Assessments. Boston: Study Group on the National Assessment of Student Achievement, 1986.

Dorr-Bremme, Don, and James Caterall. Test Use Project: Costs of Testing. Los Angeles: UCLA Center for the Study of Evaluation, November 1982.

Education Commission of the States. "National Efforts." State Education Leader, 11:1 (spring 1992).

Education Week. "By All Measures: The Debate Over Standards and Assessments." Education Week, Special Report, June 17, 1992.

Haladyna, Thomas M., Susan Bobbit Nolen, and Nancy S. Haas. "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution." Educational Researcher, 20:5 (1991).

Jaeger, Richard M. "Legislative Perspectives on Statewide Testing: Goals, Hopes, and Desires." Phi Delta Kappan, November 1991.

Koretz, Dan, et al. The Reliability of Scores From the 1992 Vermont Portfolio Assessment Program: Interim Report. Technical Report No. 355. Los Angeles: UCLA Center for the Study of Evaluation, December 1992.

Madaus, George F. "The Effects of Important Tests on Students: Implications for a National Examination System." Phi Delta Kappan, November 1991.

Madaus, George F., 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.

McLaughlin, Milbrey W. "Test-Based Accountability as a Reform Strategy." Phi Delta Kappan, November 1991.

McRae, Douglas J. "TOPIC: Too Much Testing?" Press release. Monterey, Calif.: CTB Macmillan/McGraw-Hill, November 15, 1990.

Mehrens, William A., S.E. Phillips, and Christine M. Schram. Survey of Test Security Practices. East Lansing, Mich.: Michigan State University, 1992.

National Association of Elementary School Principals. "Standardized Tests Useful—But Don't Need More, Say Principals." Press release. Alexandria, Va.: March 27, 1992.

National Commission on Testing and Public Policy. From Gatekeeper to Gateway: Transforming Testing in America. Boston: Boston College, 1990.

National Council on Education Standards and Testing. Raising Standards for American Education. Washington, D.C.: January 1992.

National Council of Teachers of Mathematics. Curriculum and Evaluation Standards for School Mathematics. Reston, Va.: 1989.

Pechman, Ellen M., and Peirce A. Hammond. A Background Report on Educational Assessment. Washington, D.C.: National Research Council of the National Academy of Science, July 1991.

Roeber, Edward D. Survey of Large-Scale Assessment Programs. Washington, D.C.: Association of State Assessment Programs, Council of Chief State School Officers, fall 1990 and spring 1991.

Shavelson, Richard J., Gail P. Baxter, and Jerry Pine. "Performance Assessments: Political Rhetoric and Measurement Reality." Educational Researcher, 21:4 (May 1992).

Shepard, Lorrie A. "Will National Tests Improve Student Learning?" Phi Delta Kappan, November 1991.

Sinclair, Beth, and Babette Gutman. A Summary of State Chapter 1 Participation and Achievement Information for 1989-90. Prepared for U.S. Department of Education, Office of Policy and Planning, 1992.

Smith, Marshall. "Policy Issues: The Systemic Character of Reform and Implications for Curriculum and Equity." Paper presented at the annual meeting of the American Educational Research Association, San Francisco: April 1992.

Toch, Thomas. "Schools for Scandal." U.S. News & World Report, April 27, 1992.

U.S. Congress. House Committee on Education and Labor. Oversight Hearing on the Report of the National Council on Education Standards and Testing, serial no. 102-105. Washington, D.C.: U.S. Government Printing Office, February 19, 1992.

U.S. Congress. Office of Technology Assessment. Testing in American Schools: Asking the Right Questions. OTA-SET-519. Washington, D.C.: U.S. Government Printing Office, February 1992.

Wigdor, Alexandra K., and Bert F. Green, eds. Performance Assessment for the Workplace (Volume I) and Technical Issues (Volume II). Washington, D.C.: Committee on the Performance of Military Personnel, National Research Council, 1991.

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20877

Orders may also be placed by calling (202) 275-6241.

United States
General Accounting Office
Washington, D.C. 20548

Official Business
Penalty for Private Use $300

First-Class Mail
Postage & Fees Paid
GAO
Permit No. G100