

**Estimating the Costs of Student Assessment
in North Carolina and Kentucky:
A State-Level Analysis**

CSE Technical Report 408

Lawrence O. Picus
CRESST/University of Southern California

February 1996

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

Copyright © 1996 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA Catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the policies of the Office of Education Research and Improvement or the U.S. Department of Education.

ESTIMATING THE COSTS OF STUDENT ASSESSMENT IN NORTH CAROLINA AND KENTUCKY: A STATE-LEVEL ANALYSIS

Lawrence O' Picus, Alisha Tralli, and Suzanne Tacheny

CRESST/University of Southern California

Introduction

Tests are being widely touted as instruments of educational reform and renewal. State minimum competency testing has increased dramatically over the last several decades from one program in 1972 to 34 programs in 1985 (Haney, Madaus, & Lyons, 1993). Many states have been shifting from traditional, multiple-choice testing formats to alternative forms of student assessment. These new assessment methods include performance assessment, authentic testing, and portfolio review. Critics of traditional, multiple-choice assessments see this testing method as part of the cause of low educational standards because multiple-choice testing reinforces rote learning rather than complex thinking and active learning (Nickerson, 1989; Wiggins, 1991). Wiggins (1992) argues that performance assessment is a better method of determining a student's higher order thinking skills than the use of traditional, multiple-choice tests. Higher order thinking is often nonalgorithmic, complex, involves the application of multiple criteria and uncertainty, and potentially has multiple solutions. Some research documents the negative effects of standardized testing on both teaching and learning (Shepard, 1991; Smith, 1991). In order to assess more sophisticated cognitive skills, many researchers promote the use of alternative assessment methods as a means of educational reform (Barone, 1991; Krechevsky, 1991; Wiggins, 1989).

Surveys show that teachers spend considerable time teaching to test objectives, test-taking skills, and even specific test items (Haney et al., 1993). For this reason, it is suggested that the test become "worth teaching to" (Wiggins, 1992). Tests may be used as instruments to improve teaching and learning through focusing instruction on areas deemed important, such as higher order thinking skills. National and state projects have been implemented towards these goals.

The New Standards Project, a joint effort of the National Center on Education and Economy and the University of Pittsburgh's Learning Research and Development Center, is a nationwide assessment project, which began in 1991 and includes 18 states and 6 school districts. This program has received a lot of attention with its focus on assessment through performance tasks, projects, and portfolios of student work. The aim of the program is to make thinking, problem solving, and communication skills "count" by creating and fostering assessments designed to elicit them (Viadero, 1994b). The National Council on Education Standards and Testing (NCEST), a commission convened at the national level, endorses an assessment system that consists of multiple methods of measuring student progress (Koretz, Madaus, Haertel, & Beaton, 1992). One objective of Goals 2000, an act legislated by Congress in 1994, is for students to "demonstrate the ability to reason, solve problems, apply knowledge, and write and communicate effectively" (National Education Goals Panel, 1994, p. 9). These aims are similar to those of many states that have implemented new assessment methods.

Over the past several years, numerous state-level programs have been implemented that reflect more "authentic" testing practices. Kentucky's assessment system was developed to "elicit authentic performances through which students can accurately demonstrate their ability to develop, understand, and use knowledge" (Foster, 1991, p. 35). The California Learning Assessment System (CLAS) included authentic, performance-based assessment methods in order to provide a more "in-depth picture" of student abilities (California Department of Education, 1993). The North Carolina Standard Course of Study (SCS), the state-adopted curriculum, includes portfolio assessment in the first and second grades, as well as open-ended testing in the later grades. Vermont has also implemented a portfolio program through voluntary participation. State education agencies are increasingly seeking alternative assessment strategies.

Despite the importance placed on assessment systems, it is surprising how small a portion of state education funding is devoted to the development, implementation and evaluation of student assessments. As Picus (1994) suggests, simple analysis of state-level expenditures for assessment programs may mask the true cost of these assessment programs, both the newer forms of assessment and more traditional multiple-choice methods. Cost information is vital to solid policy making. Administrators and policy makers at the state and

district level need this information in order to assess the viability of implementing new testing practices. In addition, other stakeholders—such as students, teachers, parents, and taxpayers—have an interest in the cost of implementing and maintaining new educational programs. Although several states have implemented new assessment programs, there has been little research on the costs of developing and implementing these new systems. The purpose of this study is to analyze the total costs of implementing alternative assessment programs at the state level. A second study currently underway takes a careful look at the costs of assessment programs at the district level.

The Conceptual Framework

The assessment of true economic costs includes estimates of the benefits foregone from the next best alternative and comparing the benefits received with the costs incurred. However, many of the benefits of educational assessment are difficult to identify, hard to measure, and not easily valued in monetary terms. In addition, knowledge of costs is hampered by the multiplicity of possible benefits and the numerous and sometimes contradictory goals of assessment systems. For instance, the goals of providing information on student performance and providing better educational experiences for students may not coincide (Snow, 1989; Taylor, 1994). This study focuses on the state-level costs of new assessment systems, where resource flows are assumed to represent the true economic cost to the state of implementing the program. This ingredient approach to program cost determination has been used in a multitude of studies in education (King, 1994; Monk, 1993; Quinn, Van Mondfrans, & Worthen, 1984; Stecher, 1995).

The conceptual framework for this study was developed by Picus (1994). He provides a comprehensive list of the elements needed to estimate the expenditures for any alternative assessment program. The list includes the direct expenditures that would be incurred in implementing any new assessment program, but does not include the estimation of opportunity costs, few of which were identified at the state level. The concept of opportunity costs is more important at the district level, and efforts to identify such costs will be described in later reports. Relying on Levin's (1983) ingredients approach, all of the individual items that are purchased as part of an assessment program must be identified and summed to provide a complete picture of the total costs of the program. Picus uses a three-dimensional matrix to help identify all of these costs.

The first dimension of the matrix relates to the *components* of the assessment program and includes such things as the development, production, administration and scoring of the test instruments. The second dimension has to do with the *level* at which the expenditures are incurred. Expenditures may be necessary at any one of a number of levels including the state, school district, school, or classroom, or even the private test market. The third dimension of this matrix deals with the specific *kinds* of items purchased for each component at each level, be it personnel, test materials, computer resources, or travel and food for training sessions.

Figure 1 is a matrix showing these three dimensions of *component*, *level*, and *kind* of expenditure. The cost of the resources needed for each of the ingredients of the program can be placed into different cells of this matrix. To get an accurate estimate of the expenditures for any assessment program, it is necessary to identify all of the components, levels, and kinds of expenditures that must be made.

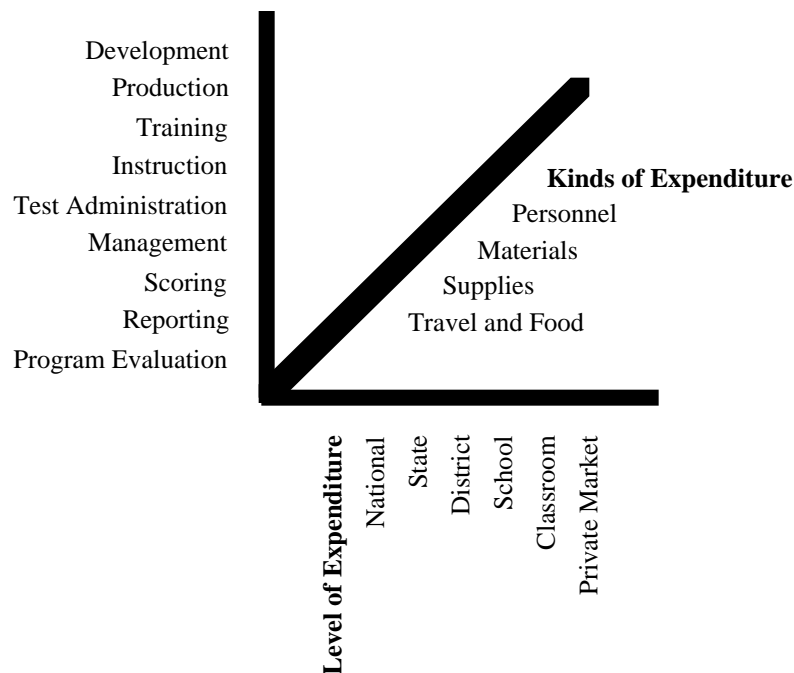


Figure 1. Expenditure dimensions.

Methodology

Definition of Terms

Costs and expenditures are not synonymous terms. Monk (1995) distinguishes between these two terms. Costs are “measures of what must be foregone to realize some benefit,” while expenditures are “measures of resource flows regardless of their consequence” (p. 365). Expenditures are generally easier to track since accounting systems typically report resource flows by object, for example, instruction, administration, transportation. The conceptual framework discussed in the prior section categorizes the assessment expenditures along three dimensions: Level of Expenditure, Kinds of Expenditure, and Components of Expenditure. This study focuses on state-level expenditures. In terms of Kinds of Expenditure, the framework has five self-explanatory categories: personnel (salary and fringe benefits), materials, supplies, travel and food, and other. Definitions for the seven Components of Expenditure categories are provided below.¹

Development expenditures are cash outflows for the purpose of developing performance assessment items. These expenditures will probably be high in the initial phase of performance assessment program development, but they can be expected to decrease as developers move up the learning curve and additional items are banked for future use. Development expenditures include all outlays for determining assessment design concept, task development, pilot studies, production, and ongoing task development. Note that this study includes production expenditures in the development expenditure category since these two resource outflows were difficult to separate as distinct categories.

Training expenditures are cash outflows for the purpose of teaching individuals to administer, score, interpret, and use new assessment practices as an instructional tool. Training expenditures are a function of the complexity of the assessment, the number of assessment items, and the location of training sessions.

Management expenditures are cash outflows associated with coordinating and overseeing the development and implementation of the assessment program.

¹ See Picus (1994) for more information on how expenditures and costs are measured.

Test administration expenditures are cash outflows for implementing the performance assessment. This includes delivery of the test and test materials to the test site, monitors and facilitators for the test, and delivery of test and test materials to the scoring site.

Scoring expenditures are cash outflows involved with scoring the performance assessment tests. Note that this does not include expenditures for development of scoring rubrics or scorer training. Development of scoring rubrics is included under the development expenditure category. Scorer training is included under the training expenditure category. Scoring expenditures are a function of the method of assessment, the number of assessees and assessors, the time for review, the complexity of the assessment instrument, and the region or population being tested.

Evaluation expenditures are cash outflows involved with assessing the strengths and weaknesses of the assessment instruments, including validity, reliability, and attainment of strategic goals.

Reporting expenditures are cash outflows involved with reporting performance assessment information. Stakeholders requiring information include students, parents, interested citizens, educators, evaluators, researchers, and administrators. Reporting costs are a function of the level of comparison (individual, school, district, state), the number of report recipients, congruence with existing reporting mechanisms, and the method of reporting results.

The state-level expenditure estimates provided in this paper for North Carolina and Kentucky focus on expenditures rather than costs. It was our evaluation of the state expenditure data that few true opportunity costs could be identified at the state level. On the other hand, the identification of opportunity costs at the district level seems more productive and more likely to yield important information about the true levels of effort devoted to alternative assessment programs across the state and their member school districts.

Data Collection

Four states were initially considered for analysis of costs related to developing and implementing alternative assessment programs: California, Kentucky, North Carolina, and Vermont. California and Kentucky were chosen because they had recently changed from more traditional testing systems to state-of-the-art alternative assessment programs. North Carolina, which has a

more traditional assessment system, provides a baseline for the costs of rigorous traditional assessment programs. Finally, Vermont is unique in instituting a portfolio assessment system where none existed previously.

California and Vermont are not included in this initial state-level analysis. The California Learning Assessment System (CLAS) was administered for only two years and was subsequently cut out of the state's budget. The tests, which were eventually to provide individual, school, and district information on student learning using multiple-choice, short-answer, essay, and performance assessment testing methods, proved to be very controversial, and funding for their continued support was vetoed by Governor Wilson in 1994. We made numerous attempts to obtain cost information from state officials. Unfortunately, our efforts corresponded with the period of greatest controversy making it impossible to meet with and collect data from officials of the California Department of Education. When funding for future years was vetoed, we felt continued efforts would be difficult and would not yield as valuable information as other potential avenues of study.

Vermont implemented a portfolio-based assessment system in the 1991-1992 school year. However, Vermont is not included in this analysis because of the voluntary nature of the assessment program and the state's small size. Our intention is to conduct both state- and district-level analyses of costs for our second report. Moreover, for purposes of assessing state-level costs, it was thought that statewide mandatory assessment systems would provide better cost estimates on our first attempt.

The information in this document on North Carolina and Kentucky was gathered through a variety of sources. Data were collected from the following individuals:

- Employees in the State Department of Education
 - Chief state school officer
 - Director of assessment program
 - Other personnel involved in the development of assessment instruction, training, scoring, dissemination, and management
- Officials of state fiscal and budget offices
- Legislative staff
 - House education committee

- Senate education committee
- Fiscal committee staff
- Legislative research or analysis staff
- Representatives of education interest groups
 - Teachers
 - Administrators
 - Board members
 - Parents
- Staff of independent task forces or commissions
- Private sector representatives, such as test publishers

Data were also collected through document review of budgets and program information, both internal and external. In Kentucky, actual expenditure data were obtained for three fiscal years, 1991-1992, 1992-1993, and 1993-1994. Because the majority of the expenditures in Kentucky are through a fixed contract, the expenditures were estimated for two additional years. In North Carolina, data were collected for a three-year period from the fiscal years 1992-1993 through 1994-1995.

Units of Analysis

Expenditure comparisons are sensitive to the unit of analysis. Typically, expenditures are evaluated in terms of per pupil measures. However, in this analysis, expenditures are broken down by three different units of comparison:

- expenditure per student enrolled,
- expenditure per student tested, and
- expenditure per test administered.

Expenditure per student enrolled is the total expenditures for the assessment program divided by the number of K–12 public school students in the state. Expenditure per student tested is the total expenditure divided by the number of students actually taking tests. Students were often given more than one test, so a final calculation gives the expenditure per test administered.

Some economies of scale can be expected. Assessment programs with more students involved will have smaller associated expenditure per student, expenditure per student tested, and expenditure per test administered because

fixed costs can be distributed over more units. In addition, those involved with the new assessment methods can be expected to become more efficient with new methodologies over time. Flexer and Gerstner (1993) found that classroom teachers became more efficient with their time devoted to performance assessment as the school year progressed.

The time period evaluated is also important. Most development costs will occur at the beginning of implementation of a new program (Monk, 1993). Although development of testing items is continuous, the up-front costs are significantly larger due to a learning curve on task and item development, as well as the need for a large number of items initially. The percentage of raw items that survive subsequent refinement and pilot testing can be expected to increase over time. As a task and item bank is filled, there is not the need to develop the large number of items required in the first years of program development and implementation.²

Limitations and Delimitations

In instances where programs that accomplish the same goals are compared, and they each have identical efficiencies in their operation, expenditures provide an adequate basis for analysis and program comparison. However, if the benefits to be derived from the two programs differ, then this approach will overstate the cost of the program with greater benefits. For instance, if one assessment program provides greater student learning, the added benefits should be subtracted from total expenditures to account for the incremental benefit when its costs are compared to other assessment programs. Similarly, a reduction of negative effects, such as misuse of student and teacher time for extraneous testing, is an additional benefit that must be accounted for. To the degree that performance assessment is superior to conventional assessment, and the degree that performance assessment replaces prevailing counterproductive practices, some portion of the expenditures is absorbed. Totaling expenditures may overstate the social costs of implementing an assessment program. This analysis does not make assumptions about factors that may cause an overstatement of

² If the assessment instruments must be made public, the development costs increase dramatically. For example, Texas has estimated that the cost of making the items used on the Texas Assessment of Academic Skills public would cost the state \$6 million a year (Clark & Picus, 1995).

costs, if they exist. Rather, expenditures at the state level are assumed to reflect state costs.

Numerous difficulties revolve around the determination of costs for educational services. Education project goals can be difficult to quantify (Haney et al., 1993), and they may be contradictory (Taylor, 1994). Opportunity costs associated with the time devoted to a new program are especially hard to determine. The opportunity cost of the additional time devoted to the new program should be determined for all participants: students, teachers, administrators, volunteers, and parents. Determining an opportunity cost for students, volunteers, and parents is especially problematic.³ However, for the purpose of determining state-level costs, expenditures are assumed to adequately represent the cost of new assessment since the opportunity costs associated with students, volunteers, and parents are primarily school-level costs.

Monk (1993) points out two other problems with the use of expenditures as a proxy for costs in performance assessment programs. First, the involvement of the state may be viewed as a serious limit on the market's ability to efficiently produce the correct mix of educational outcomes. The state provision of resources to testing and assessment may not be the same as that which the free market would apply. Expenditures on performance assessment may not be aligned with social priorities and may reflect inefficiency in the use of public resources. Second, costs will be absorbed to the degree that performance assessment reform programs substitute for other educational practices. Economic costs may be overstated by the degree to which substitution effects occur in local staff development, use of classroom time for performance assessment, and the utilization of assessment information.⁴

³ Haney, Madaus, and Lyons (1993) calculate the cost of student time devoted to testing and test preparation using net present value analysis. They use the net present value of foregone wages as the measure of opportunity cost of student time. Using foregone wages and net present value analysis is fraught with uncertainties, including the correct wage to be attached to student time, and the appropriate discount rate to be applied to the future cash flows. Similar problems occur with valuing volunteer and parent time.

⁴ In-service training is generally provided to classroom teachers prior to and during the school year. New assessment training could become part of ongoing in-service programs. With respect to classroom instruction, performance assessment can be integrated more easily into instruction, as opposed to being solely a testing function. To the extent that performance assessment provides good information to teachers and students, time may be saved from pointless testing practices.

Stecher (1995) reports that the costs of hands-on science assessments in which students construct, observe, measure, manipulate, and otherwise interact with objects and equipment are considerably higher than the costs of written constructed-response assessments. The added cost is primarily due to the additional equipment and materials needed for scientific experimentation. Therefore, the expenditures reported in this study do not necessarily reflect the resources needed to develop and implement assessment programs in other subject areas.

In summary, these expenditure numbers may be difficult to generalize to other state assessment programs. Numerous variables preclude quick comparisons. States differ in number of students, current testing systems in place, and curriculum guidelines. In addition, the number of grades tested, the number and types of subjects tested, the types of assessment, and the information desired will all impact costs and expenditures for implementing new assessment programs.

Program Background

Kentucky and North Carolina both have implemented new comprehensive assessment systems. The Kentucky system is more radical in that it clearly departs from the sole use of multiple-choice and short-answer testing methods. North Carolina has opted for a more traditional assessment system, although the testing program includes open-ended questions. Table 1 compares the Kentucky and North Carolina assessment systems.

North Carolina

In July 1993, the North Carolina legislature created a commission, the North Carolina Education Standards and Accountability Commission, to define educational standards for public schools. The role of the commission included developing fair and valid assessments. One of the commission's recommendations was the use of more authentic assessments (North Carolina Education Standards and Accountability Commission, 1994).

Table 1

Description of Performance Assessment Systems

	Kentucky KIRIS	North Carolina End of Grade/Course Tests																								
Introduction of new system	The Kentucky school system was ruled unconstitutional in 1989. The Education Reform Act (KERA) was introduced in 1991, including KIRIS. Interim testing in 1991-1992 established baseline data.	The North Carolina testing system was developed in conjunction with the state's standard course of study and expanded in 1992-1993 to include End-of-Grade and End-of-Course tests.																								
Parts of tests	<ul style="list-style-type: none"> • Transitional tests (multiple-choice and short-answer) • Performance tasks • Portfolios 	<ul style="list-style-type: none"> • Multiple-choice • Open-ended questions requiring descriptions, analysis, and written explanation 																								
Performance levels	<ul style="list-style-type: none"> • Novice (beginner) • Apprentice • Proficient • Distinguished 	<ul style="list-style-type: none"> • Level I (not passing) • Level II • Level III • Level IV (superior) 																								
Grade levels tested	Mandatory testing is given in the 4th, 8th, and 11th (changed from 12th in 1994) grades. Schools are encouraged to use old versions of the tests as practice in the other grades.	All grade levels are tested. Grades 1 and 2 use portfolios. Grades 3 through 8 use End-of-Grade tests, and Grades 9 through 12 use End-of-Course tests.																								
Number of students tested	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Tested</th> <th style="text-align: center;">ADA</th> </tr> </thead> <tbody> <tr> <td>1991-92:</td> <td style="text-align: center;">131,250</td> <td style="text-align: center;">622,089</td> </tr> <tr> <td>1992-93:</td> <td style="text-align: center;">137,452</td> <td style="text-align: center;">628,775</td> </tr> <tr> <td>1993-94:</td> <td style="text-align: center;">137,172</td> <td style="text-align: center;">627,296</td> </tr> </tbody> </table>		Tested	ADA	1991-92:	131,250	622,089	1992-93:	137,452	628,775	1993-94:	137,172	627,296	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Tested</th> <th style="text-align: center;">ADA</th> </tr> </thead> <tbody> <tr> <td>1992-93:</td> <td style="text-align: center;">1,215,714</td> <td style="text-align: center;">1,081,559</td> </tr> <tr> <td>1993-94:</td> <td style="text-align: center;">1,283,357</td> <td style="text-align: center;">1,088,998</td> </tr> <tr> <td>1994-95:</td> <td style="text-align: center;">1,298,071</td> <td style="text-align: center;">1,101,483</td> </tr> </tbody> </table>		Tested	ADA	1992-93:	1,215,714	1,081,559	1993-94:	1,283,357	1,088,998	1994-95:	1,298,071	1,101,483
	Tested	ADA																								
1991-92:	131,250	622,089																								
1992-93:	137,452	628,775																								
1993-94:	137,172	627,296																								
	Tested	ADA																								
1992-93:	1,215,714	1,081,559																								
1993-94:	1,283,357	1,088,998																								
1994-95:	1,298,071	1,101,483																								
Exclusion allowances	Exceptions to testing are made for foreign students who have spent less than two years in an English-speaking school. Special education students receive a modified portfolio.	Students in programs for "exceptional students" are administered tests that are based on individual education plans. Others may be exempt based on State Board of Education exclusion guidelines.																								
Scoring	A team of 82 scorers was temporarily hired by the Kentucky Department of Education. Preference was given to teachers.	Multiple-choice tests are scanned locally. The short-answer questions are scored at a central location by over 1,000 teachers during a three-week period.																								

Table 1 (continued)

	Kentucky KIRIS	North Carolina End of Grade/Course Tests
School and district indicators for accountability	<p>Each school has a “threshold” and a “baseline” for test results:</p> <ul style="list-style-type: none"> • > 1% above the threshold receives a financial reward; • below the threshold but above the baseline, then the school develops an improvement plan; • from 0 to 5% below the baseline, then a distinguished educator is assigned to the school to develop a plan; and • > 5% below the threshold, then the school is placed on probation. <p>Noncognitive indices include rate of attendance, drop-out rates, physical and mental health barriers, and proportion of successful postgraduate transitions.</p>	<p>Performance indicators include:</p> <ul style="list-style-type: none"> • attendance and dropouts; • “completer index”; and • academic performance indicators in all areas tested. <p>Low-performance schools are identified and action is taken when schools are in serious trouble. Scores are shown as a comparison to the state. Schools are also compared to schools with similar advantagement levels. An advantagement index is created for each district and ranges from negative 40 to positive 40.</p>
Reporting	Results are reported at the student, district, and school level. Student-level reports include parent letters.	State legislature requires annual report cards showing all schools.
Assessments replaced	CTBS/Benchmark 4 Version Tested grades were 3, 5, 7 and 10	California Achievement Test (1985-1992)
Other assessment systems currently used	<ul style="list-style-type: none"> • National Assessment of Educational Progress (NAEP) • PSAT • SAT 	<ul style="list-style-type: none"> • Minimum Skills Diagnostic Tests • North Carolina Competency Tests • National Assessment of Educational Progress (NAEP) • PSAT • SAT
Test development	Advanced Systems in Management and Evaluation (ASME), New Hampshire	Mostly managed by the North Carolina Department of Public Instruction, with assistance from North Carolina State University.

North Carolina’s current assessment system was developed in conjunction with the state’s Standard Course of Study (SCS), a state-adopted curriculum that defines what students are expected to know and do at each grade level. The new tests were developed and revised by North Carolina teachers, curriculum

specialists, and test development consultants. The new tests “reflect higher standards for student performance” and “place increased emphasis on higher level thinking and application skills” (North Carolina Department of Public Instruction, 1992a, pp. 1-2). The SCS provides a foundation to develop and improve instruction tests.

In 1992-1993, North Carolina expanded the testing system to include End-of-Grade (EOG) and End-of-Course (EOC) assessments. Students in the first and second grades compile a portfolio of work, which parents and teachers can both review to determine student progress towards SCS goals. Third- through eighth-grade students are given EOG tests, and ninth- through twelfth-grade students are given EOC tests in selected subject areas. EOG tests are administered at the end of the school year to test mastery of grade-level knowledge and skills. When fully implemented, five curricular areas will have EOG tests: reading, mathematics, science, social studies, and writing. At the time of this study only reading, writing, and mathematics EOG tests were being implemented. EOC tests are administered at the end of each course; however, not all subject area tests are currently available.⁵

Training expenditures can be divided into three categories: (a) training individuals to administer the assessment; (b) training individuals to score, interpret, and use test results; and (c) training individuals in related instructional strategies linked to the assessment goals.

Management of test development and implementation was through the North Carolina Department of Public Instruction.

EOG tests are administered at different times. The tests in reading and mathematics are given in Grades 3 through 8 at the end of the school year.⁶ The writing test is administered in February for students in Grades 4, 6, and 8. EOC tests are administered to students who are enrolled in the specified courses for graduation credit at the end of the course term.

⁵ In 1992, tests were being administered in algebra I and II; geometry; biology; physical science; physics; chemistry; U.S. history; economic, legal, and political systems; and English I and II (North Carolina Department of Instruction, 1992b). Tests in additional course areas will include English III, world studies, earth sciences, and other areas as funding and resources become available.

⁶ The science and social studies EOG tests will also be administered to the same six grade levels at the end of the school year when they are developed.

The tests in North Carolina include multiple-choice and short-answer questions. The advantages of this system are easier scoring, quick feedback, reliability, and validity. The multiple-choice portions of the tests are scored at each school district using microcomputers provided by the state. Each district received from one to seven of these computers, and a similar number of college student interns. The purpose of the computers is to enable same- or next-day scoring of the multiple-choice portion of the EOC and EOG tests. The college students work in the school districts during the summer helping central office staff, site administrators and classroom teachers analyze and interpret the results of the tests.

The open-ended items are scored almost as quickly. Immediately following the end of school in June, some 1,000 teachers are hired to come to the state capital in Raleigh to spend one week scoring the open-ended tests. Each teacher spends one week working in a university gym. Training takes place on Monday morning, and the teachers spend the rest of the week scoring the test instruments using pen lights and bar codes to enter the results into computers. Staff analyze the scores immediately for reliability, and offer teachers who slip from the standard short retraining sessions through the week. Test results are available to schools and districts almost immediately at the end of the three-week scoring sessions.

The North Carolina Department of Instruction plans to develop a way to establish reference points on the EOG tests that relate to performance on other nationally published standardized tests. By using these reference points, students can be categorized within bands of national achievement (North Carolina Department of Public Instruction, 1992a). However, this type of evaluation has not been implemented at this time. Current evaluation places students in one of four categories.

There are four categories of performance (North Carolina Department of Public Instruction, n.d.). Students performing at Level I “do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.” Students performing at Level II “demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.” Students performing at Level III “consistently demonstrate mastery of grade-level subject matter and skills and are well

prepared for the next grade level.” Students performing at Level IV “perform in a superior manner clearly beyond that required to be proficient at grade-level work.”

Annual reporting on all schools is required by the state legislature. Individual scores for multiple-choice tests are available to students, teachers, and parents within a few weeks after testing. School-system scores are made available but take longer to compile and report. Both EOG and EOC test results from all school systems are published in state reports in the first quarter of the year that follows testing.

Kentucky

Kentucky’s approach to assessment includes a combination of transitional multiple-choice and short-answer tests, which provide validity and stability, as well as performance assessments and portfolio requirements. It is hoped that the new authentic tests will promote better instructional practices as teachers teach to the new tests (Foster, 1991). Noncognitive indicators are also used to assess schools and districts.⁷ In the 1991-1992 school year, an interim testing program was administered that provided baseline data for determining future performance. Testing is mandatory in the fourth, eighth, and twelfth grades. However, legislation in 1994 moved the twelfth-grade performance assessment to the eleventh grade.

The assessment system currently used in Kentucky is one of nine components in a massive, systemwide program to reinvent the state’s educational system. The Kentucky Supreme Court decision in *Rose v. Council for Better Education* in 1989 overturned the state’s entire educational system and resulted in the state’s legislature passing the Kentucky Education Reform Act of 1990 (KERA). KERA was organized around three major themes: curriculum, governance, and finance. KERA contains nine specific objectives for improvement (Legislative Research Commission, 1990): (a) assessment, rewards, and sanctions; (b) equalization of expenditures and prohibition of nepotism; (c) family resource and youth service centers; (d) preschool programs; (e) primary school program; (f) extended school services; (g) school-based decision making; (h) professional development; and (i) technology in education. The first goal resulted in the creation of the Kentucky Instructional Results Information System (KIRIS)

⁷ The following noncognitive indicators have been identified as potentially informative (Kentucky Instructional Results Information System, 1993): attendance; drop out; retention; physical and mental barriers to learning; and transition to work, postsecondary education, and the military.

to develop a performance-based assessment program. The new assessment system includes portfolios in mathematics and writing, and testing that involves the application of knowledge to real-life problems. Testing is mandated in the fourth, eighth, and eleventh grades.

Test development committees were formed in each content area—reading, mathematics, science, and social studies. The committees were made up of classroom teachers, resource teachers, consultants, administrators, curriculum coordinators, university professors, Kentucky education association representatives, and Kentucky Department of Education personnel. The framework was taken from both the Kentucky learning goals and the National Assessment of Educational Progress (NAEP). The new assessment system was field tested. The basic elements of all of the test questions are developed by teachers and then fine-tuned and field tested by the staff of an outside contractor.

Kentucky has contracted with Advanced Systems in Management and Evaluation (ASME), a New Hampshire firm, to develop and manage the new statewide assessment system. The \$29.5 million, five-year contract was the largest of its kind in the United States (Viadero, 1994a). ASME has provided the bulk of the development responsibilities, although the Kentucky Department of Education provided some management resources, primarily in the form of personnel time. Development expenditures include both the resources for initial design and implementation of the assessment program and expenditures for the continuous development of assessment items.

Training goals of test administration, scoring, interpretation, and use of assessment instruments in related instructional strategies overlapped in Kentucky. In particular, training in how to score the assessment instrument had a great deal of overlap with training in instructional strategies related to the new program. The training to implement the new assessment system was pyramidal. Each regional coordinator trained fifteen to twenty cluster leaders in his or her region. The cluster leaders, in turn, trained the teachers at the school level.

Management of the new assessment program was primarily a function of Kentucky's Department of Education, which oversaw the progress of ASME. However, ASME also had internal management functions.

Test administration, which includes the costs of delivery of materials both to and from school sites, as well as the personnel costs associated with test security,

was the responsibility of ASME. Student portfolios are evaluated by teachers at the school site. ASME periodically checks the results for accuracy and reliability. ASME provides all personnel, equipment, and supplies necessary to score the performance assessments, multiple-choice tests, and short-essay questions.

Students are classified into one of four performance levels. There are four categories of performance (Kentucky Department of Education, n.d.). Students performing at Level I “do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.” Students performing at Level II “demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.” Students performing at Level III “consistently demonstrate mastery of grade-level subject matter and skills and are well prepared for the next grade level.” Students performing at Level IV “perform in a superior manner clearly beyond that required to be proficient at grade-level work.”

The assessments are not used to make student-level decisions, although students learn their scores. The test results are used to assess how well schools are doing. Assessment results are used to guide the system of rewards and sanctions at the school level. The state is undertaking two independent evaluations of the new assessment system. One is by the legislature’s Office of Accountability and the other is by Western Michigan University and the Kentucky Institute for Education Research in Frankfurt.

Assessment results are reported at the student, school, and district levels. Student reports include item-level reports for multiple-choice items in each of the four content areas, item-level reports for performance-based tasks, student summary reports, and parent letters. Initial reports were not made on the writing portfolio performance. At the school and district levels, a single student assessment curriculum report is prepared with numbers and percentages of students at each performance level in each curricular area.

Cost Analysis

The assessment systems of North Carolina and Kentucky are different in many respects. First, North Carolina uses more traditional testing methods than Kentucky, which has elected to use more authentic assessment methodologies. Second, the level of accountability is different for the two states. In North

Carolina, testing provides data at the student level, while in Kentucky, the data are primarily used for school-level decision making. For these reasons, comparisons between the states must be made cautiously. The analysis below separates the data for the two states.

North Carolina

The expenditures associated with North Carolina’s assessment program are summarized by components in Table 2. Changes in the assessment system that were begun in the 1992-1993 school year were managed internally by existing personnel. In North Carolina, state-level expenditures associated with training are relatively small.

Table 3 presents the expenditure information in a percentage format. As mentioned above, there is little training involved in either administering or scoring the assessments, so training expenditures are negligible. Test administration took over half of the resources devoted to the new assessment system. Test administration includes test delivery expenditures and personnel expenditures. Individuals, whether from the district or the school level, who assist in the administration of the assessment instrument are a resource outflow.

Table 2
North Carolina State-Level Component Expenditures in Dollars (Thousands)

Component	1992-1993	1993-1994	1994-1995	3-Year total
Development	\$1,199	\$1,400	\$1,400	\$3,999
Training	\$0	\$0	\$0	\$0
Management	\$1,355	\$1,355	\$1,355	\$4,064
Test administration	\$4,206	\$5,409	\$5,370	\$14,985
Scoring	\$2,216	\$597	\$859	\$3,672
Evaluation	\$1	\$75	\$75	\$151
Reporting	<u>\$61</u>	<u>\$292</u>	<u>\$298</u>	<u>\$651</u>
Totals	\$9,038	\$9,127	\$9,357	\$27,522

Table 3

North Carolina State-Level Component Expenditures (Percentage)

Component	1992-1993	1993-1994	1994-1995	3-Year total
Development	13%	15%	15%	15%
Training	0%	0%	0%	0%
Management	15%	15%	14%	15%
Test administration	47%	59%	57%	54%
Scoring	25%	7%	9%	13%
Evaluation	0%	1%	1%	1%
Reporting	<u>1%</u>	<u>3%</u>	<u>3%</u>	<u>2%</u>
Totals	100%	100%	100%	100%

Table 4 summarizes the assessment expenditures per student enrolled, per student tested, and per test administered. Virtually all students were tested with an average expenditure per test of \$7.25 per student tested. On average, students took 1.67 tests, that is, many students took more than one test; and the average expenditure per test administered over the three-year period was \$4.59. As Table 4 indicates, enrollment, number of students tested, number of tests administered, and total assessment expenditures in North Carolina have been fairly stable over

Table 4

North Carolina Expenditure per Student Enrolled, Expenditure per Student Tested, and Expenditure per Test Administered

Component	1992-1993	1993-1994	1994-1995	3-Year total
Average daily attendance (ADA)	1,081,559	1,088,998	1,088,998	3,259,555
Number of students tested	1,215,714	1,283,357	1,298,071	3,797,142
Total tests administered	2,000,000	2,000,000	2,000,000	6,000,000
Total assessment expenditure	\$9,037,599	\$9,127,359	\$9,356,859	\$27,521,817
Expenditure per ADA	\$8.36	\$8.38	\$8.59	\$8.44
Expenditure per student tested	\$7.43	\$7.11	\$7.21	\$7.25
Expenditure per test administered	\$4.52	\$4.56	\$4.68	\$4.59

Note. North Carolina makes every possible effort to test all students; therefore the number of students tested does in fact exceed the average daily enrollment.

the three-year analysis period. This has led to relatively consistent expenditures per student enrolled, per student tested, and per test administered.

North Carolina has spent over \$9 million per year developing and implementing a new assessment system. The new system has administered 6.0 million tests to 3.8 million students. Expenditures have remained consistent over the time period studied yielding an average expenditure of \$4.59 per test administered. The state costs for this assessment system represent approximately one quarter of one percent (0.26%) of the total state expenditures for K–12 education in North Carolina each year, and 0.17% of total expenditures for K–12 education in that state.

Kentucky

Kentucky signed a five-year contract with ASME to develop and implement a new assessment program. The budget for the first year, 1991-1992, was largely used to develop and introduce the new assessment program. The budget for 1992-1993 was largely for interim testing of the system. The budget for 1993-1994 provides the budget for the fully operational program. For the first three years, Table 5 provides actual expenditures for development and implementation of Kentucky's new assessment system by component. These expenditures reflect both the contract costs for ASME and resource outflows by the Kentucky Department of Education. Virtually all of the management expenditures are from Kentucky's Department of Education, while the rest of the costs are ASME contract costs. While the Department of Education did not hire additional personnel to perform the various management tasks related to the new assessment program, an expenditure is associated with the time that employees redirect to the new assessment function. ASME expenditures were 74% of expenditures in 1991-1992, 80% of expenditures in 1992-1993, and 83% of expenditures in 1993-1994. Conversely, Kentucky Department of Education personnel time and materials dedicated to the new assessment program accounted for 26%, 20%, and 17% of resource outflows over this time period. The largest component of Kentucky Department of Education expenditures was for employee salaries and fringe benefits, which accounted for over 80% of their expenditures in the first three years.

Table 5

Kentucky State-Level Component Expenditures in Dollars (Thousands)

Component	1991- 1992	1992- 1993	1993- 1994	1994- 1995	1995- 1996	5-Year total
Development	\$1,247	\$1,855	\$2,020	\$2,158	\$2,271	\$9,551
Training	\$216	\$391	\$925	\$990	\$1,042	\$3,564
Management	\$1,394	\$1,382	\$1,513	\$150	\$235	\$4,674
Test administration	\$501	\$654	\$713	\$763	\$804	\$3,436
Scoring	\$895	\$1,580	\$2,233	\$2,387	\$2,513	\$9,608
Evaluation	\$25	\$27	\$30	\$32	\$0	\$114
Reporting	<u>\$424</u>	<u>\$436</u>	<u>\$476</u>	<u>\$509</u>	<u>\$536</u>	<u>\$2,380</u>
Totals	\$4,703	\$6,325	\$7,910	\$6,989	\$7,401	\$33,327

The expenditure figures in Table 5 for 1994-1995 and 1995-1996 reflect ASME contract expenditures and estimated management expenditures for the Kentucky Department of Education. Kentucky Department of Education costs are based on analysis of the Department's budget and discussions with staff as to the portion of responsibilities devoted to the KIRIS system. As in North Carolina, the state-level assessment costs represent a very small portion of total state K-12 education expenditures. For example, in 1993-94, the \$7.9 million spent on assessment represented less than one-half of one percent (0.45%) of state K-12 education expenditures, and 0.28% of total school district revenue.

Table 6 lists each component of expenditure as a percentage of expenditures by year and as a percentage of total expenditures for the five-year period.

Table 6

Kentucky Performance Assessment State-Level Component Expenditures (Percentage)

Component	1991- 1992	1992- 1993	1993- 1994	1994- 1995	1995- 1996	5-Year total
Development	27%	29%	26%	31%	31%	29%
Training	5%	6%	12%	14%	14%	11%
Management	30%	22%	19%	2%	3%	14%
Test administration	11%	10%	9%	11%	11%	10%
Scoring	19%	25%	28%	34%	34%	29%
Evaluation	1%	0%	0%	0%	0%	0%
Reporting	<u>9%</u>	<u>7%</u>	<u>6%</u>	<u>7%</u>	<u>7%</u>	<u>7%</u>
Totals	100%	100%	100%	100%	100%	100%

Table 7 provides an analysis of the assessment expenditures per student enrolled, student tested and test administered. The cost per student tested is substantial, a result of the multiple tests administered through the system. Unlike North Carolina, which tests every student, Kentucky relies on a sample of students.

Summary of Analysis

Koretz et al. (1992) estimate that Advanced Placement exams, which are similar to some types of performance assessment, cost \$65 per subject test, while commercial standardized tests cost from \$2 to \$5 per subject test. Stecher (1995) reports that the cost of the complete five-hour CTBS battery of multiple-choice tests is \$2.80 per student, with the cost increasing significantly when open-ended written responses are included in the assessment. He reports that the CTB writing test costs \$4.80 per student for a single prompt. We estimated that the North Carolina system, which included a combination of both multiple-choice and open-ended questions, cost just over \$4.50 per test administered.

Stecher (1995) also analyzed the cost of performance assessment in science, and found the per student cost of one performance task ranged from \$17 to \$85, depending on whether best-case or worst-case assumptions were used, as well as the number of students tested. In addition, he reports that approximately four tasks would be needed to produce a student score with a reliability comparable to the ITBS Science Subtest. Our estimates of the Kentucky testing system showed that costs ranged from almost \$6.00 per test administered to approximately

Table 7

Kentucky Expenditure Per Student Enrolled, Expenditure Per Student Tested, and Expenditure Per Test Administered

Component	1991-1992	1992-1993	1993-1994	3-Year total
Average daily attendance (ADA)	569,713	582,054	594,859	1,746,626
Number of students tested	131,250	140,000	143,100	414,350
Total tests administered	787,500	840,000	892,500	2,520,000
Total assessment expenditure	\$4,702,561	\$6,325,094	\$7,909,988	\$18,937,643
Expenditure per ADA	\$8.25	\$10.87	\$13.30	\$10.84
Expenditure per student tested	\$35.83	\$45.18	\$55.28	\$45.70
Expenditure per test administered	\$5.97	\$7.53	\$8.86	\$7.51

\$8.86 per test administered over the three years for which data are currently available, somewhat lower than Stecher's estimates of science performance assessments. However this comparison must be made in light of the fact that Kentucky's assessment instruments are designed to test multiple subjects and not just science. If performance assessments for science are more costly due to added equipment and materials needed, then our findings are not surprising when compared to Stecher's. Finally, the cost analysis in this report includes only those costs identified at the state level, and additional costs may be expected at the district and school levels.

Conclusions

With fiscal retrenchment and recession affecting many states and local economies, policy makers will increasingly be concerned with costs associated with new educational programs. Although many of the assessments are being touted as new alternatives to traditional multiple-choice tests, testing via essay and oral examination has a much longer history than multiple-choice examinations, which were a creation of the twentieth century because of the expense of using and the difficulties in standardizing these kinds of assessment when used with large numbers of people (Haney et al., 1993). Godfrey and Conboy (1994) state that knowledge of costs of activities, including hidden costs, is necessary for cost control.

It is much more problematic to compare cost estimates for different types of pupil assessment programs than to determine the cost for a single program. Such comparisons require controls for differences in the nature and magnitude of the benefits being generated. As this study focused on expenditures, a comparison between the North Carolina and Kentucky programs should not be made. In addition, these two programs were implemented with different purposes in mind, and therefore the costs and outcomes are expectedly different.

In comparison with more traditional multiple-choice examinations, performance assessment and portfolio requirements are time-intensive since fewer tasks can be administered in a single time period, and they take longer to score and evaluate. Fewer tasks give a smaller base on which to judge score reliability and validity (Burger, 1994; Mehrens, 1992; Viadero, 1994b). The cost of implementing new testing methodologies will depend on the level to which the state wants accountability.

For instance, Vermont's portfolio assessment system was not designed to provide information at the individual level; rather, the information is collected and reported at the school and district levels. On the other hand, California's CLAS assessment was intended to eventually provide information at the individual level, which requires much more testing information. Likewise, the North Carolina testing was developed to provide student-level information, while the Kentucky assessment was developed primarily to provide information at the school and district levels for accountability purposes.

Although innovative types of performance-based assessments focus on higher order and complex cognitive skills, researchers caution that cost, practicality, comparability, generalizability, objectivity, and administrative convenience should be considered before development and implementation of a program (Koretz et al., 1992). The numbers presented in this paper provide expenditure estimates for specific programs implemented in two states. Generalizations should be made cautiously, as economies of scale, level of information desired, testing systems in place, and testing methodologies will all affect the cost and expenditures of developing and implementing new assessment programs at the state level. What is surprising, given the tremendous emphasis placed on assessment systems to measure school accountability, is the relatively minuscule portion of educational expenditures devoted to this important, and highly visible, component of the educational system.

References

- Barone, T. (1991). Assessment as theater: Staging an exposition. *Educational Leadership*, 48(5), 57-59.
- Burger, S. E. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice*, 13(1), 9-15.
- California Department of Education. (1993). *Students, standards and success: The California Learning Assessment System*. Sacramento, CA: California Department of Education.
- Clark, C., & Picus, L. O. (1995). *Preliminary report on CPRE core study for Texas* (Typescript). New Brunswick, NJ: Consortium for Policy Research in Education.
- Flexer, R. J., & Gerstner, E. A. (1993). *Dilemmas and issues for teachers developing performance assessments in mathematics* (CSE Tech. Rep. No. 364). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.
- Foster, J. D. (1991). The role of accountability in Kentucky's Education Reform Act of 1990. *Educational Leadership*, 48(5), 34-36.
- Godfrey, H., & Conboy, R. M. (1994). Cost control in North Carolina public schools. *Government Finance Review*, 10(4), 27-30.
- Haney, W. H., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Boston, MA: Kluwer Academic Publishers.
- Kentucky Department of Education. (n.d.). *Quick guide to student assessment in Kentucky*. Frankfort, KY: Kentucky Department of Education.
- Kentucky Instructional Results Information System. (1993). *1991-92 technical report*. Frankfort, KY: Kentucky Instructional Results Information System.
- King, J. A. (1994). Meeting the educational needs of at-risk students: A cost analysis of three models. *Educational Evaluation and Policy Analysis*, 16(1), 1-19.
- Koretz, D. M., Madaus, G. F., Haertel, E., & Beaton, A. E. (1992). *National educational standards and testing: A response to the recommendations of the National Council on Education Standards and Testing*. Santa Monica, CA: RAND.
- Krechevsky, M. (1991). Project Spectrum: An innovative assessment alternative. *Educational Leadership*, 48(5), 43-48.

- Legislative Research Commission, (1990). *A guide to the Kentucky Education Reform Act of 1990*. Frankfort, KY: Legislative Research Commission.
- Levin, H. (1983). *Cost analysis: A primer*. Thousand Oaks, CA: Sage.
- Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, 11(1), 3-9, 20.
- Monk, D. H. (1993). *The costs of systemic education reform: Conceptual issues and preliminary estimates*. Ithaca, NY: Cornell University, Department of Education.
- Monk, D. H. (1995). The costs of pupil performance assessment: A summary report. *Journal of Education Finance*, 20(4), 363-371.
- National Education Goals Panel. (1994). *Data volume for the national education goals report, volume one*. Washington, DC: U.S. Government Printing Office.
- Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher*, 18(9), 3-7.
- North Carolina Department of Public Instruction. (n.d.). *The new tests: More challenging for students*. Raleigh, NC: Author.
- North Carolina Department of Public Instruction. (1992a). *Getting to know the new tests: Questions and answers*. Raleigh, NC: Author.
- North Carolina Department of Public Instruction. (1992b). *Quick reference for parents and teachers: Testing at various grade levels*. Raleigh, NC: Author.
- North Carolina Education Standards and Accountability Commission. (1994). *Initial report to the North Carolina State Board of Education, the North Carolina General Assembly, and Governor James B. Hunt, Jr.* Raleigh, NC: Author.
- Picus, L. O. (1994). *A conceptual framework for analyzing the costs of alternative assessment* (CSE Tech. Rep. No. 384). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.
- Quinn, B., Van Mondfrans, A., & Worthen, B. R. (1984). Cost-effectiveness of two math programs as moderated by pupil SES. *Educational Evaluation and Policy Analysis*, 6(1), 39-52.
- Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, 73, 232-238.
- Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5), 8-11.
- Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher*, 18(9), 8-14.

- Stecher, B. (1995). *The cost of performance assessment in science: The RAND perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal*, 31, 231-262.
- Viadero, D. (1994a). The little firm that could. *Education Week*, June 8, pp. 25-27.
- Viadero, D. (1994b). Teaching to the test. *Education Week*, July 13, pp. 21-25.
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 703-713.
- Wiggins, G. (1991). Standards, not standardization: Evoking quality student work. *Educational Researcher*, 48(5), 18-25.
- Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, 49(8), 26-33.