A New Conceptual Framework for Analyzing the Costs of Performance Assessment

Lawrence O. Picus
Frank Adamson
William Montague
Margaret Owens

scope
Stanford Center for
Opportunity Policy in Education

This study was conducted by the Stanford Center for Opportunity Policy in Education (SCOPE) with support from the Ford Foundation and the Nellie Mae Education Foundation.

The Stanford Center for Opportunity Policy in Education (SCOPE) supports cross-disciplinary research, policy analysis, and practice that address issues of educational opportunity, access, equity, and diversity in the United States and internationally.

Citation: Picus, L. O., Adamson, F., Montague, W., & Owens, M. (2010). *A New Conceptual Framework for Analyzing the Costs of Performance Assessment*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu
http://edpolicy.stanford.edu

scope
Stanford Center for
Opportunity Policy in Education

Table of Contents

Preface and Acknowledgements

This paper is one of eight written through a Stanford University project aimed at summarizing research and lessons learned regarding the development, implementation, consequences, and costs of performance assessments. The project was led by Linda Darling-Hammond, Charles E. Ducommun Professor of Education at Stanford University, with assistance from Frank Adamson and Susan Shultz at Stanford. It was funded by the Ford Foundation and the Nellie Mae Education Foundation and guided by an advisory board of education researchers, practitioners, and policy analysts, ably chaired by Richard Shavelson, one of the nation's leading experts on performance assessment. The board shaped the specifications for commissioned papers and reviewed these papers upon their completion. Members of the advisory board include:

Eva Baker, Professor, UCLA, and Director of the Center for Research on Evaluation, Standards, and Student Testing

Christopher Cross, Chairman, Cross & Joftus, LLC

Nicholas Donahue, President and CEO, Nellie Mae Education Foundation, and former State Superintendent, New Hampshire

Michael Feuer, Executive Director, Division of Behavioral and Social Sciences and Education in the National Research Council (NRC) of the National Academies

Edward Haertel, Jacks Family Professor of Education, Stanford University

Jack Jennings, President and CEO, Center on Education Policy

Peter McWalters, Strategic Initiative Director, Education Workforce, Council of Chief States School Officers (CCSSO) and former State Superintendent, Rhode Island

Richard Shavelson, Margaret Jacks Professor of Education and Psychology, Stanford University

Lorrie Shepard, Dean, School of Education, University of Colorado at Boulder

Guillermo Solano-Flores, Professor of Education, University of Colorado at Boulder

Brenda Welburn, Executive Director, National Association of State Boards of Education

Gene Wilhoit, Executive Director, Council of Chief States School Officers

The papers listed below examine experiences with and lessons from large-scale performance assessment in the United States and abroad, including technical advances, feasibility issues, policy implications, uses with English language learners, and costs.

- Jamal Abedi, *Performance Assessments for English Language Learners.*

- Linda Darling-Hammond, with Laura Wentworth, *Benchmarking Learning Systems: Student Performance Assessment in International Context.*

- Suzanne Lane, *Performance Assessment: The State of the Art.*

- Raymond Pecheone and Stuart Kahl, *Developing Performance Assessments: Lessons from the United States.*

- Lawrence Picus, Frank Adamson, Will Montague, and Maggie Owens, *A New Conceptual Framework for Analyzing the Costs of Performance Assessment.*

- Brian Stecher, *Performance Assessment in an Era of Standards-Based Educational Accountability.*

- Barry Topol, John Olson, and Edward Roeber, *The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments.*

An overview of all these papers has also been written and is available in electronic and print format:

- Linda Darling-Hammond and Frank Adamson, *Beyond Basic skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning.*

All reports can be downloaded from http://edpolicy.stanford.edu.

We are grateful to the funders, the Advisory Board, and these authors for their careful analyses and wisdom. These papers were ably ushered into production by Barbara McKenna. Without their efforts, this project would not have come to come to fruition.

Introduction

The use of assessments in schools continues to expand alongside our growing expectations of what school children need to know and be able to do when they graduate. The standards-based approach to education and education reform has focused attention on the role of standardized tests and created higher stakes for both students and the education professionals who serve them. At the same time, the press for continued improvement in student learning—as measured by state established learning and performance standards—has caused schools to rely more heavily on assessment tools as part of "data-driven decision" making processes. Despite the growing importance of assessments in our education system, relatively little is known about the economic costs and benefits of these assessments that are such a large part of every student's educational experience.

What is clear, however, is that following the passage of the No Child Left Behind (NCLB) Act in 2001 and the requirement that states measure the progress students make toward meeting student proficiency goals, the amount of, and level of testing in schools has increased dramatically—along with the costs for those assessments. The U.S. Government Accountability Office[1] ([GAO], 2009) estimated that, in the 40 states responding to its survey, payments to testing vendors to develop, administer, score, and report results of assessments to meet NCLB requirements exceeded $640 million in 2007–2008. This figure does not include the costs of NCLB testing in the other 10 states or the costs of tests developed and administered by any of the 50 states themselves. Some estimates place the total costs of NCLB-required testing in the United States at $1 billion annually (GAO, 2003), with costs much greater for test formats that include more open-ended items and performance tasks.

Because testing costs are closely related to the kinds of items and tasks used—and because those tasks significantly influence the cognitive demands of the tests and their potential influence on instruction—it is important to evaluate the costs and benefits of performance assessments. This paper updates earlier work on estimating the costs of alternative assessments within the current policy context (Picus, 1994; Picus, Tralli, & Tasheny, 1996; Picus & Tralli, 1998), with a focus on the costs of developing, administering, scoring, and reporting the results of performance assessments. As in the earlier studies, efforts are made to distinguish between the concept of economic or opportunity costs (i.e., the use of teacher time that is already "paid for" through the contract and used as part of the assessment process rather then for some other activity or function), and the direct expenditures made for assessment.

As the paper shows, the bulk of the costs for any assessment system are the time teachers and other school and district personnel spend in the development, administration,

grading, and use of the results of assessments, not the costs of the assessment activities themselves. Similarly, and not unexpectedly, the benefits of assessments depend on the extent to which those same individuals are able to use the data from those assessments to improve student learning or performance.

It is relatively straightforward to determine how much a school or school district spends on assessment instruments and reporting of the test results, but it is much more difficult to determine how much time is devoted to preparing for and administering the tests, and even harder to determine the costs of how the results of those assessments are used by school staff to improve learning and instruction. These estimates are made more complex by the growing realization that teacher collaboration through Professional Learning Communities (DuFour, DuFour, Eaker, & Many, 2006) or similar teacher efforts are critical to improving student learning.

Much of the literature on improving student performance describes collaborative, data-driven approaches that rely heavily on analysis and use of student assessments. These efforts are time-consuming on the part of school staff and often require extensive training to be fully implemented. When used as part of a strategy to improve student learning, there are multiple benefits to these systems as well. Knowing how performance assessments can lead to higher student performance through better identification of student needs and more appropriate approaches to teaching is a critical component of analyzing the costs of assessment systems.

One of the concerns frequently raised about performance assessments is the cost of scoring more extended, open-ended items in relation to the costs of machine-scored multiple-choice tests. However, many states and nations have maintained performance assessment systems that are manageable and affordable. At least two aspects of this problem need to be explored and estimated in cost/benefit terms: first, the manner in which the assessments are (or are not) integrated into teachers' work—with scoring managed as part of teachers' professional development (which has both cost and benefit implications)—and second, the extent to which benefits of teachers' participation in this work translates into improved instruction and learning for students.

Another issue for consideration when estimating the costs and benefits of performance assessment is the identification of benefits. The traditional literature on cost/benefit analysis (see for example, Mishan & Quah, 2007) focuses mostly on the monetization of costs and benefits. While this analysis establishes a framework for (and some initial estimates of) the costs assessment practices, the benefits of assessment are measured in terms of student performance, which is not easily translated into dollars. Cost- effectiveness analysis (Levin & McEwan, 2002) offers an alternative, which is considered in this paper.

This paper focuses on the use of assessments to improve learning and offers a framework for estimating some of the costs and benefits of performance assessments, includ-

ing the influences on both costs and benefits of different scoring models (for example, widespread training and involvement of teachers as part of their ongoing work and professional development; external scorers unrelated to the classroom; uses of technology platforms for facilitating scoring). Following our development of this conceptual framework, we provide recent estimates of assessment costs.

This paper has five sections. It starts with a framework for considering assessments by establishing working definitions of formative, benchmark, and summative assessments as they will be used in this paper. The second section advances earlier work on the concept of costs of assessments compared to direct expenditures for assessments in schools. The third section focuses on the benefits in terms of improved student performance. Section 3 is followed by a table (Table 3) that summarizes the costs, expenditures, and benefits of various types of assessments. The fourth section offers an appraisal of what we currently know about expenditures for assessment and testing today. Finally, the fifth section summarizes the framework developed and offers suggestions for improving the analysis of the costs and benefits of performance assessments.

Framework for Categorizing Assessments

In today's standards-based environment, assessment of student performance is an accepted and regular part of the expectations of all schools. Students are familiar with the annual tradition of "standardized" tests—generally required by each state and given in the spring of each school year to measure how students and schools perform. Parents, the media, and even real estate agents, eagerly await the reporting of these test data to "see how well their schools are doing." In many states, one can visit a web site to compare school-level test results (along with related student demographic information) across schools and school districts. Some states, such as California, have even reduced the reporting of test results (along with other measures of a school's "success") into a single index number. Moreover, the press by schools and districts to make Annual Yearly Progress (AYP) to avoid sanctions identified in the federal NCLB law has all schools working to help students do better on those standardized tests.

Yet, as argued below, state-wide standardized tests are only one part of the entire student performance assessment system available for use by schools. Many argue that to fully understand student needs, provide instructional programs to meet those needs and assess the effectiveness of the programs, a tiered structure of assessment is needed. (See, for example, Boudett, City, & Murnane, 2008). If one is to identify the costs of assessment programs, these distinctions among what is known as formative, benchmark, and summative assessment become important. Unfortunately, educators do not have a clear agreement regarding the distinction across these three levels of assessment. Thus, before estimating costs and benefits, it is important to establish a framework for types of assessment.

As used in this paper, formative assessments are diagnostic and include teacher-developed tools to understand what students know and need to know. Benchmark assessments are periodic tests to check understanding to ensure students have mastered the material they have been taught. Summative assessments are the standardized annual tests given in virtually every school. These summative assessments are often used to measure school success or quality. Boudett, City, and Murnane (2008) also look at these as short, medium, and long-term data, but their distinctions are similar to the way formative, benchmark, and summative are used below.

Formative Assessments

Odden (2009) describes formative assessment as being diagnostic in nature and given with relative frequency—sometimes as often as weekly or even daily. These assessments are used by teachers to determine how to teach specific curriculum units and to monitor regular student progress. Boudett, City, and Murnane (2008)

point out that short- term data can be generated continuously as teachers use students' regular work (assignments and tests) to assess their progress, diagnose problems with understanding, and tailor instruction to focus on areas where students need additional help or focus. This effort goes beyond the simple grading of papers, quizzes, and tests, and requires teachers to link the students' work with the learning goals of each unit, through examination of that work, observation of student participation, and conferring with students on a regular basis (Boudett, City, & Murnane, 2008).

The information derived from formative assessments is not always easy to translate into instructional practice and can require considerable work on the part of a teacher. Wylie and Lyon (2009) suggest substantial professional development (PD) efforts are needed to ensure that teachers can develop, use, and take advantage of formative assessment processes. They argue that formative assessments require school-based PD for teachers supported by coherent district support for PD efforts.

The advantage of a strong formative assessment is that it allows teachers to focus instructional activities to the exact learning status or needs of students in their classroom. Odden (2009) states that strong formative assessments that allow teachers to emphasize what students need to learn and move more quickly over material students have mastered could be thought of as being more "efficient," a concept critical to the analysis of costs and benefits.

Benchmark Assessments

Teachers and schools need periodic assessments of student progress and learning. As used in this paper, benchmark assessments provide these guideposts to educators so they can measure student progress more frequently than the once-a-year standardized state tests. The purpose of these assessments is to track student progress during a school year and might include commercially available tests or even locally developed instruments (Boudett, City, & Murnane, 2008). The value of benchmark assessment tools is that they give routine and regular progress reports on student learning to teachers and enable them to adjust their teaching strategies and pacing to ensure students are mastering the material. This is distinct from a formative assessment that helps a teacher understand what students already know, and instead gives regular and periodic information on what students have actually learned from the material presented.

Where material in one semester builds on material learned the previous semester—or any time block that is relevant to the subject matter being taught—benchmark tests enable educators to know if students are prepared or not for the new material. If not, reteaching of that material may be more efficient then the development of interventions for students who are unable to keep up—a potential benefit that could be ascribed to assessments.

Summative or Long-Term Assessments

Summative tests can include any measures that are used to make an assessment of a student's knowledge and skills at a moment in time for the purpose of drawing inferences about his or her achievement and for informing decisions. Today, most discussions of student performance and most measures of school success focus on annual statewide standardized test scores, although these are not the only summative tests in use. Statewide test data, which are often used for accountability at district and state levels, provide a snapshot of a school's performance. They can be used by schools to focus on areas needing improvement over time, and can give a good picture of gaps in achievement among different groups of students in a school or district. Unfortunately, these tests are often given in the spring with results provided in the summer or following fall, limiting their use to focus on the learning needs of students.

Summary

This section of the paper establishes a framework for various forms of performance assessment using these familiar terms: formative, benchmark, and summative assessment. As used here, formative assessments are often highly individualized by teacher, used frequently and in various forms to identify what students need to learn to master the material being taught. Formative information can also be secured from large-scale assessments if they are sufficiently rich and if data are delivered to teachers in a detailed and timely way.

Benchmark assessments are given on a regular and periodic basis throughout the school year and are designed to measure how well students have learned the material presented. They allow teachers to make corrections in their teaching to ensure that students have the knowledge from early units needed to master the more difficult skills and knowledge required in higher units of the curriculum. Summative assessments in today's U.S. policy system are typically annual standardized tests that allow for schools to see how well they are meeting state-established standards over time as well as how they compare with other schools in the district and state.

While there are many ways to draw distinctions among formative, benchmark, and summative assessments, what is important here is establishing definitions to facilitate the design of the cost framework provided in the next section.

Measuring the Costs of Assessment

Before developing a conceptual framework for measuring the costs of performance assessment, it is important to establish the difference between the concepts of costs and expenditures. Each is discussed briefly below.

Expenditures

A common approach to comparing the costs of alternative programs in educational institutions is to determine the monetary value of the resources necessary to implement each program, and compare the total expenditures across programs. Economists point out that this process implicitly assumes the two programs are intended to accomplish the same goals, and that both have identical efficiencies or inefficiencies in their operation. If these conditions do not hold, and there is little reason to expect that they do, comparisons of expenditures are invalid and can be misleading (Monk, 1990; Belfield, 2000).

If, as is often the case in education, there are multiple goals established for an alternative assessment program, then estimation of the costs of that program must include all of the resources necessary to accomplish all of those goals. The difficulty is that a project's goals can be hard to quantify or may even be contradictory. For example, among the many goals that have been attributed to performance assessment are: to change what is taught and learned in schools focusing more on problem solving and critical thinking; to raise expectations of students; and to motivate student interest and effort in learning. Determining the resources necessary to achieve each of these goals is, at best, a complex task. Because of this difficulty, many analysts stop short of estimating the true costs of a program, and instead focus on the expenditures required for its implementation.

In K–12 educational institutions, even determining the actual expenditures for a specific program can be difficult. Most state-accounting systems require school districts to report spending by object (salaries, benefits, supplies, etc.), and sometimes by function (instruction, administration, instructional support, maintenance and operations, transportation, etc.). Odden and Picus (2008) point out that often these expenditure data are reported at the district level, and there is little or no information about how funds are used at the school or classroom level. Moreover, detailed information about specific programs within a district is often hard to discern from school district financial reports. In an object-oriented system, estimating the expenditures for student assessment might require determining the salaries and benefits of staff members who work in that program, estimating what portion of their time is devoted to the assessment program, and then determining which of the district's expenditures for supplies and materials (including the tests) should be attributed to the program. These expenditures may be coded in different places in the district's accounting reports, making their estimation more difficult (Hartman, 2002).

Even in districts able to provide detailed information about the expenditures made for their assessment program, this information only provides a partial delineation of the full economic costs of the assessment program. The other factors that must be considered when estimating the full costs of a program are described below.

Costs

The textbook definition of the cost of a program is the benefits that are not realized through the best forgone alternative. Thus, if a resource is devoted to some use, the benefits associated with the best possible alternative use of that resource represent the "opportunity cost" of the program. Unfortunately, it is not always possible to determine what the best alternative use of those resources might be. Moreover, if that alternative can be identified, determining its benefits may be a considerable problem. For example, if a district is considering the implementation of a new performance assessment program, the opportunity costs of that program would be equal to the benefits from any conceivable alternative reform that was not implemented.

In analyzing the costs of performance assessments, the range of alternative programs could be thought of as all the possible alternative programs the district could establish to improve student performance. In this case, the benefits derived from the performance assessment would be compared to the benefits derived from the best option facing the district other than the assessment program. The more beneficial the alternative given up, the more it will cost to devote resources to performance assessment (Monk, 1995). However, before the benefits of a program can be measured, agreement must be reached as to the goals of the forgone activity.

In some cases, it may be appropriate to restrict the alternatives considered. For example, when analyzing the costs of an assessment program, it may well be that the decision to be made is whether or not to replace the existing conventional assessment system with a new form of assessment. In that case, the relative set of alternatives is limited to the assessment program currently in place, and the costs of the new assessment program will be measured on the basis of the forgone benefits of the old assessment program.

It is also unlikely that the new assessment program will require exactly the same level of resources consumed by the old system, and it is also possible that a district would be reluctant to eliminate its entire previous assessment program and shift entirely to a new system overnight. In both cases, the total resources devoted to assessment would need to be increased. In the first case, if the new program replaced the old program in its entirety, but required more resources than the old program, the forgone benefits would include both the forgone benefits of the

old assessment program plus all the benefits forgone from other activities due to the transfer of resources to the assessment program.

Similarly, if the two programs were operated together, the costs of the performance assessment would include the benefits forgone by shifting resources into assessment from other areas. If the old assessment program were only continued partially, then the costs of the performance assessment program would include the portion of the benefits forgone from the old program, along with any other benefits forgone through the resources that were shifted to the assessment program.

To make a cost analysis useful to decision makers, cost analysts need to develop a common metric to measure the benefits of alternatives. Unfortunately, there is no simple way to compare the benefits of programs that have disparate goals. Since agreement on the spectrum of benefits may also be difficult to achieve, and since estimation of the benefits of a forgone alternative may require a great deal of time for what could be considered an activity with little value (after all, why calculate the benefits of something you do not plan to do?), many analysts simplify the issue by estimating the expenditures necessary to operate the alternative program. One approach is to use the dollar value of the actual or anticipated expenditures as a measure of the projects costs. Often called the ingredients method (Levin & McEwan, 2002), this approach relies exclusively on expenditures to measure costs, and as Monk (1995) argues, leads to confusion about the difference between expenditures and costs.

If one believes that the benefits to be derived from an alternative assessment dramatically exceed the system being replaced, or if one anticipates improvements in student learning as a result of the new assessment system (clearly a hoped-for outcome of today's assessment spectrum), then using the expenditures devoted to the performance- assessment program may, in fact, overstate the true costs of the program since the benefits derived exceed the benefits from the program or programs it replaces. Unfortunately, there is no way to estimate the size of this exaggeration. To resolve this problem, one needs to make explicit assumptions about what factors could cause this overstatement, and then estimate costs with and without an adjustment for this issue. In the framework established below, the ingredients approach to costs is used, and where necessary, adjustments for the potential overstatement of benefits that this could lead to are identified and possible adjustments considered.

Establishing a Framework for Identification of Costs and Expenditures

In earlier work on the issue of assessment costs (Picus, 1994; Picus, Tralli, & Tasheny, 1996; Picus & Tralli, 1998), Picus proposed three dimensions of costs or expenditures for assessments. Picus identified them as components, kinds, and levels. The factors identified in each category are outlined in Table 1 (page 10).

Table 1. Dimensions of Costs/Expenditures for Performance Assessments

Dimensions		
Kind	Component	Level
Personnel	Development	National
Materials	Production	State
Supplies	Training	District
Travel and Food	Instruction	School
	Test Administration	Classroom
	Management	Private Market
	Scoring	
	Reporting	
	Program evaluation	

The costs/expenditures identified in Table 1 might be thought of as a three- dimensional matrix whereby costs/expenditures could be located in any cell that related to each of the three dimensions. For example, personnel costs/expenditures are likely to be incurred at all levels (national, etc.) and for most of the components of assessment (management and scoring, for example).

The basic cost/expenditure factors identified in Table 1 have not changed noticeably in the past 15 years, but the relative allocation of each may have changed considerably. For example, the availability of online testing capacity such as Northwest Evaluation Association's (NWEA's) Management of Academic Progress (MAP) test have changed the way many assessments are now scored, enabling teachers to have the results of these benchmark tests the next day—data they can use to adjust instruction as they move forward with lesson plans. Also, the adaptive approach used in many of the computer- based assessment systems—assessments that adjust the questions asked of students based on their responses—provide much more useful and accurate measures of student knowledge and skills.

At the same time, the press for frequent formative and even benchmark assessments may lead to the use of more of the resource factors identified in Table 1 than has been the case in the past. The important question that policy makers, school district administrators, school site leaders, teachers and even parents need to address is the extent to which the use of multiple assessment strategies offers benefits in terms of enhanced or improved student achievement. Given the multiple variables that impact the administration of such assessments (identified below), it is likely impossible to assign direct benefits to the cost/expenditures of assessment programs.

However, substantial evidence suggests that when including well-thought-out student assessment programs in an overall school reform strategy, they are part of an effective program that leads to improved student performance (Odden & Picus, 2008; Odden & Archibald, 2009; Odden, 2009). Table 2 below repeats the components identified in the center column of Table 1 but consider whether they should be thought of as "costs," "expenditures," or both, and identifies the potential benefits within each component. In addition, Table 2–4 look at these components across the three types of assessment identified above—formative (Table 2), benchmark (Table 3), and summative (Table 4).

The data in Table 2–4are intended to represent the types of costs, expenditures, and benefits that schools, districts, and states would encounter in the development, implementation, and analysis of performance-assessment programs. These tables provide a framework for thinking about the costs of assessment systems. What is clear from Table 2–4 is that most of the costs relate to personnel time to be allocated in different ways. Thus, there may be little change in personnel expenditures, and a change in the focus of what teachers, counselors, site leaders, and others at a school site are doing. The actual expenditures for materials and supplies are likely relatively low.

Addressing the Benefits of Performance Assessment

Thinking about the benefits of performance assessment is more complex. While no one doubts the value of assessing student performance, there is a growing movement among educators who argue that time spent in assessment activities is taking away from the time needed to teach material to the students. It is hard to evaluate the accuracy of statements like this because the findings would be dependent on the quality of the assessment and the way it is used. Several ways to think about the benefits (positive and negative) of assessment are presented below:

- If properly aligned with state or local standards, assessments are a useful tool to help assess student progress and learning needs, and can help schools and districts identify strengths and weaknesses.

- Standardized tests only provide data for one point in time, and the results are often not available until the next year when a student has moved to another classroom and often another school. Nor do standardized tests measure progress when there is a high incidence of student mobility.

- Benchmark assessments will give schools and teaches regular information on student progress during the school year and help them focus.

- Well-designed formative assessments can be used as tools to sharpen instruction to focus on student needs leading to improved student outcomes over time.

- Assessment results can form the basis of teacher collaboration and planning efforts to design instruction programs that are coherent and focused directly on student learning needs, leading to improved student outcomes.

- A strong assessment system will enable teachers to focus on student learning problems earlier, resulting in fewer expensive interventions.

- Assessment takes time away from instruction and limits student learning.

- Standardized assessments are often misused and are not representative of what is taught in classes, resulting in little value for teachers or students.

The problem with all of these potential benefits is the multiplicity of outcomes to measure and the fact that so much of the benefit to be derived is dependent on the way the assessments are implemented, analyzed, and then used to drive and, hopefully, improve instruction. Thus, unless assessment is an integral component of an overall school strategy to improve learning, it is unlikely that any assessment program will provide a high level of benefits by itself. This, of course, complicates the measurement of those benefits.

For example, in our recent work in Wyoming and Little Rock, we are finding that much of the instruction in classrooms seems to spend relatively little time in initial instruction, with more time focused on providing interventions for students when they fall behind. In most of the schools we have visited, specific times during the day are set aside for all students to have interventions, under the assumption that they will need additional help. This time is provided at the expense of instructional time in the core classes (math, science, language arts, social studies, and world languages).

Better formative assessments, combined with a focus on high-quality instruction to start with would, it seems, reduce the need for expensive interventions. To the extent that well-designed formative and benchmark assessments are used to identify and resolve student learning concerns early and quickly, parsing out the benefits that can be attributed to the assessment is probably impossible—yet it is clear that if well-designed, a model like this could result in more high-quality instruction for all students and fewer but more timely interventions for students when they do struggle with the material.

The benefits of performance assessment then seem to fall into several categories. They provide more information to help teachers identify and correct learning deficiencies early. They give teachers and site leaders information about how well their students are learning the material required by a state's standards, and they provide long-term information on changes in overall student performance across schools and districts. All of these can be used to focus and design instruction to improve student learning, and are clearly a benefit of performance assessment.

Table 2. A Framework for Measuring the Costs and Expenditures of Performance Assessment—Formative Assessment

(Bold text represents more important cost/expenditure components, while unbolded text represents lesser impact on overall costs.)

Formative Factors	Costs	Expenditures	Benefits
Development	These costs are largely for local teacher time both to develop instruments and establish a system to monitor student progress. This would be classified as a cost, not an expenditure, because the teacher time is already paid for through the contract.	There may be a few expenditures for materials and supplies beyond what would be utilized otherwise, but these seem minor.	The benefits from development of assessments (formative, benchmark, or summative) would come from insights about how to assess student learning more effectively that help teachers improve instruction and student learning, measuring such benefits in terms of dollars is probably impossible.
Production		These are simply the costs of duplicating any tests/quizzes, etc., the teacher chooses to use. They could be somewhat higher if more sophisticated materials are used.	
Training	Odden and Picus (2008) offer a PD model that includes 10 days of pupil-free PD for teachers—which could include time for development and use of formative assessments. The model also includes the extensive use of coaches who could help teachers analyze formative assessment results. To the extent district budgets include these resources for PD, this would represent a cost; to the extent that additional PD time for teachers (additional days in the contract, for example) or additional staff to serve in the roll of coaches, there would be additional expenditures. Teachers also need time for collaboration, either paid for as part of the pupil-free PD days, or as part of the existing teacher contract through planning/ collaboration time during the school day.	Expenditures would accrue in instances where additional days are added to teacher contracts for PD, for substitute time during the school year, and to pay trainers who work with teachers on issues of assessment. The complexity of estimating these expenditures results from separating expenditures explicitly for assessment from other PD expenditures.	The benefits from the training component accrue from additional knowledge, skills, and teaching strategies learned by the teachers as part of the PD program. Time spent with coaches in activities that improve instruction could also provide benefits in terms of student learning.
Instruction	The concept here is not teaching students or teachers, but rather adjusting instruction based on the results of the formative assessment. Under that assumption, this effort would fall under the concept of a cost as teachers would use their preparation time to modify instruction and teaching based on the results of the assessments.	Theoretically, there are no additional expenditures assuming teachers use the same instructional time, only to do different things based on the information gleaned from formative assessments.	The benefits from these costs accrue through improved student learning, as well as enhanced teacher skills and hopefully improved teacher satisfaction with their work.

(continued on next page)

Table 2 (cont'd)

Formative Factors	Costs	Expenditures	Benefits
Test Administration	This is mostly a cost as the assessments would take place during class time and thus replace other activities in the classroom. Expenditures are likely to be minimal as most formative assessments could be done as part of the classroom activities.	If a school or district elects to use commercially available formative assessments such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) or materials from the Wireless Generation Website, then there may be some actual expenditure required.	
Management	These costs would be mostly transferring teacher and site leader activities from other activities to the design, implementation, and analysis of formative assessment results, and would be paid for through the existing contracts.		Potential benefits come from improved operation of a school resulting from better communication among staff and between site leadership and teachers.
Scoring	This is mostly time teachers spend scoring and analyzing the outcomes of the assessments, and would link closely with the instruction category above where the adjustments in instructional activities/strategies would take place.	To the extent that individuals are paid extra for work scoring exams (i.e., writing exams), there would be direct expenditures in this category as well.	Benefits are the added knowledge that teachers gain about student knowledge and skills enabling them to focus instruction on areas of need making classroom learning time more effective and efficient.
Reporting	Since formative assessments are mainly for the use of individual teachers, these costs would be minimal, and really part of a teacher's contracted time on the job.	Expenditures would only be the costs of materials used in conducting the assessment.	Benefits are similar to those identified above for scoring specifically the knowledge that teachers gain about student knowledge and skills enabling them to focus instruction on areas of need making classroom learning time more effective and efficient. Another potential benefit of formative assessments would be the ability to meet student needs during initial doses of instruction, rather than reverting to substantial interventions to ensure the student masters important material.
Program Evaluation	This represents the time teachers spend evaluating the helpfulness of the formative assessments in designing instruction to help improve student learning. Based on how well the instruments used by the teacher worked in helping them better understand student needs, formative assessment instruments would be modified to improve their ability to predict student needs. The later would be a benefit as well as a "cost" based on the time they spent on the task.		Based on how well the instruments used by the teacher worked in helping them better understand student needs, formative assessment instruments would be modified to improve their ability to predict student needs. The later would be a benefit as well as a "cost" based on the time they spent on the task.

Table 3. A Framework for Measuring the Costs and Expenditures of Performance Assessment—Benchmark Assessment

(Bold text represents more important cost/expenditure components, while normal text represents lesser impact on overall costs.)

Benchmark Factors	Costs	Expenditures	Benefits
Development	Benchmark assessments are generally commercial products. In this instance it is assumed that private companies that develop these assessment instruments will capitalize development costs into the price of the assessment to districts. In this analysis, focus is on the district/school costs, which are thus assumed to be mostly the time devoted to choosing specific assessment instruments. If a district develops its own benchmark assessment instruments [and Boudett, City, and Murnane (2008) describe the challenges for doing this], then there are likely to be substantial personnel costs associated with the development of the benchmark tests.	Expenditures relate to the type and frequency of assessments as well as to the "scale" of the assessment used. Districts that develop their own benchmark-assessment instruments are likely to spend more per pupil for benchmark assessments than are those districts or schools that use existing benchmark assessments developed by consortia, a state, group of states, or a national provider.	Consistent data across a school/district regarding student learning offers benefits in terms of the design of effective instruction as well as enabling teachers and site leadership to assess both student performance and, if possible, teacher performance as well.
Production	If developed in-house, the costs would be time taken away from other learning and school/district management activities. If developed commercially, the costs would be absorbed by the publisher, and passed on to districts and schools that purchase the assessments.	These expenditures would be absorbed by the publisher and shared across all clients if purchased from a commercial vendor, but would likely be substantial if developed in-house.	
Training	To properly administer, evaluate, and use the results of any benchmark assessment, teachers will require time (and probably the help of a trainer—at least initially) to learn how to use the benchmark exams to better improve student performance.	Expenditures would accrue in instances where additional days are added to teacher contracts for PD, for substitute time during the school year, and to pay trainers who work with teachers on issues of assessment. The complexity of estimating these expenditures results from separating expenditures explicitly for assessment from other PD expenditures.	These costs could also be a benefit if, as suggested above, the training leads to improved teaching and increased access to good instruction for students.

(continued on next page)

Table 3 (cont'd)

Benchmark Factors	Costs	Expenditures	Benefits
Instruction	As above, this represents time for adjusting instruction based on the results of the benchmark assessment.		And as above, benefits accrue through improved student learning, as well as enhanced teacher skills and hopefully improved teacher satisfaction with their work.
Test Administration	Costs and expenditures will depend on the form of the assessment, but the time invested in the testing is a cost, while the material (paper, pencil, test booklets, etc., or online access) represent expenditures for the administration of the assessment.	Costs and expenditures will depend on the form of the assessment, but the time invested in the testing is a cost, while the material (paper, pencil, test booklets, etc., or online access) represent expenditures for the administration of the assessment.	
Management	This is likely a cost in terms of the time of teachers, site, and district leaders to ensure the assessment is administered as intended when intended.	Expenditures would accrue for district level assessment and evaluation staff, particularly if the benchmark assessments required additional staff, and at the school site if there is an assessment coordinator who is paid or receives some form of teaching relief for his/her work as the assessment coordinator.	
Scoring	Depending on the form of the assessment, this could be time- consuming and represent either a cost of personnel time (as it is in lieu of other things they might do) or an expenditure to pay teachers (or others) to do the scoring. The costs will vary considerably depending on if the test is online, or uses Scantron type forms, or if the assessment includes writing samples that must be individually read and graded with appropriate checks for reliability and consistency in the scoring.	Depending on the form of the assessment, this could be time- consuming and represent either a cost of personnel time (as it is in lieu of other things they might do) or an expenditure to pay teachers (or others) to do the scoring. The costs will vary considerably depending on if the test is online, or uses Scantron type forms, or if the assessment includes writing samples that must be individually read and graded with appropriate checks for reliability and consistency in the scoring.	Benefits accrue to the extent that the results of the assessments are used to evaluate instructional practice and then used to improve teaching and instruction and to evaluate curriculum.
Reporting	The costs and expenditures will vary with the extent to which the results of benchmark assessments are distributed. If the findings stay within a school, reporting costs are mostly minor production costs and the time it takes to explain the meaning of the results to each teacher, and as appropriate to students and their parents.	The costs and expenditures will vary with the extent to which the results of benchmark assessments are distributed. If the findings stay within a school, reporting costs are mostly minor production costs and the time it takes to explain the meaning of the results to each teacher, and as appropriate to students and their parents.	Benefits result from discussion of testing outcomes and how students, parents, teachers, and site leaders use the data to improve teaching and learning.

(continued on next page)

Table 3 (cont'd)

Benchmark Factors	Costs	Expenditures	Benefits
Program Evaluation	As with any program, evaluation of its value in meeting goals is critical. It will take time of leadership and teachers to review the impact that continued benchmark assessments will offer in terms of focusing instruction in ways that lead to improved student performance.	If a formal evaluation of alternative benchmark-assessment instruments were conducted, there would be expenditure either for a third-party evaluation or expenditures for on-site personnel charged with evaluation responsibilities. Less formal evaluations by district/site personnel more likely would be opportunity costs for the time spent on the evaluation.	Benefits are improved assessments and better understanding of how assessment data can lead to improved instruction.

Table 4. A Framework for Measuring the Costs and Expenditures of Performance Assessment—Summative Assessment

(Bold text represents more important cost/expenditure components, while normal text represents lesser impact on overall costs.)

Summative Factors	Costs	Expenditures	Benefits
Development	If districts participate in helping assessment companies or states develop tests by piloting test items and providing review of items and test procedures without compensation from the developer (a likely outcome for state developed tests), these would be opportunity costs facing the school or district.	Summative, or standardized assessments, are generally commercial products, or instruments developed in an individual state. It is assumed that private companies that develop these assessment instruments will capitalize development costs into the price of the assessment to districts. In states that develop their own testing systems, the costs of development must be considered, as well as the ongoing costs of replacing test items to keep the instruments valid and reliable.	To the extent that schools/districts participate in the development of assessments, they will benefit from enhanced knowledge about what is tested and how it relates to state standards.
Production		To the extent that a standardized test is a paper-and-pencil exam, there are substantial expenditures associated with the production of text booklets, proctor instructions, and packaging of the tests and materials for the return of the completed test forms and booklets.	
Training	Costs are associated with the time devoted to learning how to administer the tests and, more importantly, the time teachers spend teaching such things as test- taking strategies.	There would be expenditures for training individuals to administer assessments.	
Instruction	The cost here is time spent on direct instruction focused on the test. To the extent that assessments are linked to state standards, this may not be a large "cost" or, more correctly, the benefits may outweigh the costs.		To the extent that assessments are linked to state standards, there may be considerable benefits as instruction focuses on what is assessed.

(continued on next page)

Table 4 (cont'd)

Summative Factors	Costs	Expenditures	Benefits
Test Administration	Costs and expenditures include time to prepare for, administer, and return the test materials and are potentially extensive, particularly if each school has an assessment coordinator who gets time relief from teaching or other responsibilities to administer the test at their school/district. If the school/district has a person who would not otherwise be on staff for this function, it would be an expenditure.	Costs and expenditures include time to prepare for, administer, and return the test materials and are potentially extensive, particularly if each school has an assessment coordinator who gets time relief from teaching or other responsibilities to administer the test at their school/district. If the school/district has a person who would not otherwise be on staff for this function, it would be an expenditure.	
Management	Similar to those identified in administration above.	Similar to those identified in administration above.	
Scoring		This is an expenditure that is either part of the contract for the test itself or incurred by the state/district to score the assessments.	
Reporting	The costs and expenditures will vary with the extent to which the results of summative assessments are distributed, although one would anticipate they are widely available at the school, district, and probably state level. Reporting costs are mostly minor production costs and the time it takes to explain the meaning of the results to each teacher and, as appropriate, to students and their parents. There may also be costs associated with the reporting of test results in the press, as school leaders need to react to the news that is published.	The costs and expenditures will vary with the extent to which the results of summative assessments are distributed, although one would anticipate they are widely available at the school, district, and probably state level.	Information on student performance will help focus instruction and learning in the school/district, and help each teacher better focus their teaching.
Program Evaluation	Mostly time to evaluate the usefulness of the test as well as to identify if the assessment is providing data that are useful to policy makers given the alignment of the assessment to state standards.		

Analyzing Expenditures for Assessment

Most work analyzing the costs of assessment in reality consider state, district, and school level expenditures for assessment programs. In this section, we identify current estimates of those expenditures.

In their 2002 work on test-based accountability, Hamilton, Stecher, and Klein (2002) point out that, while improved testing systems are likely to cost more money, few good estimates of the costs of improved accountability systems in relation to their benefits have been developed. Seven years later, little has changed. Few good estimates of the costs of current assessment systems exist, with much less data on what it would cost to develop and administer performance assessments.

GAO Cost Analyses

The GAO has conducted a number of analyses of assessment and testing studies over time. In 1993, a GAO study to estimate the cost of a national assessment included two components: purchase cost and time cost. The study defined purchase cost as the money spent on test-related goods and services, a category in line with what we call expenditures (U.S. GAO, 1993). The GAO also estimated the cost in terms of the time teachers, administrators, and other school personnel spent on all test-related activities, including "developing the test; preparing students to take the test; getting trained to administer the test; administering the test; collecting, sorting, and mailing completed tests; scoring the tests; and analyzing and reporting the results" (U.S. GAO, 1993). The GAO then converted the cost in terms of time into a dollar amount by multiplying the total time spent on test-related activities by the average salary in each district. Unfortunately, aggregating these different types of time disguises important differences between them that, in fairness to the GAO, have emerged in the NCLB era as more important considerations than in previous decades. Specifically, test-preparation time for students has become a subject of national debate about how much class time teachers spend "teaching to the test."

In its analysis, the GAO does provide aggregate time estimates. However, it does not provide disaggregated estimates of teacher time, nor estimated benefits in terms of either teacher PD or improved student learning. Placing a per-capita dollar amount on these benefits is complex and would mostly require extensive additional study. However, the GAO's work does point out the importance of assessing the value of time expenditures in determining the costs and benefits of assessments.

The performance assessments studied by the GAO also do not demonstrate much variety. Most included only writing samples, reading comprehension and response, and math/science problem-solving items. A few districts used science lab work, group work, and skills observations, but most still relied on paper-and-pencil testing (U.S. GAO,

1993). Of the districts and states the GAO surveyed, 30% of all tests contained some performance element, but 40% of those were writing samples alone or test batteries that included a writing sample (U.S. GAO, 1993). Only 18% of all tests asked students to perform in more than one subject area using performance-assessment formats (U.S. GAO, 1993). In every instance, test developers crafting the performance-based tests started from scratch, writing test questions that fit the state's curriculum or guidelines, then testing the draft on pilot groups of students and using an iterative revision process that did not involve state curriculum, which was undergoing simultaneous development (U.S. GAO, 1993). The 10 tests required an average of 14 months to develop from initiation to the pilot-test stage and 27 months to final form (U.S. GAO, 1993).

When reporting the cost, the GAO provided several estimates based on the type of items that policymakers might choose to include in a test. Using data from the 1990–1991 school year, the GAO proposed that an assessment including performance-based items in addition to multiple-choice items would cost about $20 per student, while a solely performance-based assessment would cost about $33 per student (U.S. GAO, 1993). Adjusting for inflation, those estimates would be closer to $32 and $53, respectively, in 2009 dollars.[1] By comparison, the GAO found that a strictly multiple-choice test would cost about $15 per student (roughly $24 in 2009 dollars). While the GAO's estimates identified a 65% larger cost for performance assessment than multiple-choice testing, $33 still only represented 0.7% of per-student expenditures in 1991 (U.S. GAO, 1993).[2]

Furthermore, the U.S. GAO (1993) made several points that highlighted potential cost-saving efficiencies. First, they reported a large spread in the cost of performance assessment, from $16 to $64 (with an average of $33). This spread suggests the potential for economies of scale and experience in developing and implementing performance assessments. When including more students in test administrations, the study found that costs fell, with fixed costs distributed over a larger number of students. In addition, when a test administration had several purposes, such as testing the same student population in more than one subject area, the per-subject-area cost of a test also declined as fixed costs were divided over a larger number of subjects. Finally, GAO researchers found performance-assessment costs to be the lowest in the states and Canadian provinces with the most years of experience administering a performance assessment, pointing towards a possible learning curve in performance-assessment efficiency (U.S. GAO, 1993). In these two regions, the cost of performance assessment averaged only $22 per student (approximately $35 in 2009 dollars).

---

1. All calculations of the cost in 2009 dollars were made using the CPI calculator available from the U.S. Bureau of Labor Statistics at http://data.bls.gov/cgi-bin/cpicalc.pl. In this paper, figures in parentheses after dollar figures represent the 2009 CPI-adjusted amounts.
2. Average expenditure per enrolled student in the U.S. was $4,902 in 1991 according to Snyder and Dillow (2010).

RAND Corporation Cost Analyses for Science

In the mid-1990s, RAND also conducted a study looking at the cost of performance assessment in science. Its researchers estimated the cost of one period of "hands-on science testing" administered to 100,000 students to be about $30 per student in 1993—or roughly $45 in 2009 dollars (Stecher, 1995). However, drawing conclusions about the cost of a large-scale, performance-assessment system on the basis of this study requires caution. RAND performed its study on a relatively small scale and tested only 2,200 students, making it difficult for the researchers to estimate the cost of any potential economies of scale (although they did find some). In addition, the small scale of the study meant that researchers did not include the costs necessary for the analysis and reporting of results that must accompany any large-scale testing system— costs which might be higher than the costs of reporting traditional standardized tests given the additional complexity of the assessment itself.

Finally, the study involved an inquiry-based science task in which students conducted an experiment and wrote a report on the results. This form of assessment, by the researchers own admission, is "among the most costly performance tests to produce because of the added expense of equipment and materials, so these estimates may represent upper bounds for the cost of performance testing of similar scope in other subjects" (Stecher, 1995).

Single- and Multiple-State Cost Assessment Analyses (Prior to NCLB)

Some cost estimates also exist from states that previously used a form of statewide performance assessment prior to the introduction of NCLB. Picus et al. (1996) conducted a study of state-level testing expenditures in Kentucky and North Carolina in the early to mid-1990s. At the time, multiple-choice questions constituted the bulk of North Carolina's assessments, though they also included a few open-ended items. The state tested students in reading and mathematics in Grades 3–8 and in writing in Grades 4, 6, and 8. At the high school level, students took a state test upon completion of certain, specified courses (Picus et al., 1996). According to Picus and his colleagues, North Carolina represented a "more traditional assessment system" than Kentucky, which employed primarily the Kentucky Instructional Results Information System (KIRIS), an innovative assessment mechanism that included portfolio and performance tasks to test students in Grades 3, 8, and 11 (Picus et al., 1996). As Kentucky shifted away from primarily multiple-choice tests to KIRIS, the state also administered a series of transitional tests that included multiple choice and short answer items alongside the more performance-oriented tasks that KIRIS required.

Picus et al. (1996) found that for the 1992–1993 through the 1994–1995 fiscal years, state-level expenditures in North Carolina averaged $4.59 ($6.51) per test administered. By comparison, for fiscal years 1991–1992 through 1993–1994, Kentucky spent an average of $7.51 ($10.65) per test administered. In the case of North Carolina, this represented 0.26% of state expenditures on K–12 education and, in the case of Kentucky,

0.45% (Picus et al., 1996). While somewhat more expensive than North Carolina's system, expenditures on KIRIS still did not amount to a large fraction of the overall per-capita state cost of education. However, Picus et al. (1996) only examined expenditures at the state level and ignored district expenditures, meaning that these figures likely underestimate actual testing expenditures in each state. Their estimation also made no effort to account for the cost of state, district, or school employees' time, as did the GAO's aforementioned 1993 study.

Picus and Tralli (1998) attempted to measure both of these missing elements (district expenditures and time spent) in their 1998 study of the cost of testing in Kentucky and Vermont. They found that the cost of testing in Kentucky—when measuring both district expenditures and the cost of school and district employees' time—were considerably higher than expenditures at the state level. Including all of the time teachers spent on KIRIS-related activities would result in a cost per test in Kentucky in 1995-1996 of between $141.41 ($200.59) and $298.66 ($423.64).

However, these estimates overstate the true cost of KIRIS, as much of the time that teachers reported spending on KIRIS-related activities was actually instructional, PD, and class-prep time. In fact, what the Picus & Tralli study counts as "costs" could also be viewed as "benefits" of the KIRIS system, as teachers needed to devote a percentage of their time to working on the portfolio and performance tasks that the system required-tasks that were designed to push students to utilize and develop higher-level thinking skills. We develop such an argument below.

Hardy (1995) presents an analysis of multiple states and their expenditures on performance assessment divided into three areas: development, administration, and scoring. Development tasks include creative and quality-control tasks leading to an assessment exercise ready for large-scale use and interpretation. Activities might include the identification and specification of the learning/assessment objectives; exercise writing; editing, review, and other quality-control procedures; small-scale pretesting; developing guidelines for scoring and interpretation; and possibly norming. When developed by an external agency, prices for performance assessments reflect these expenditures. However, in-house development by current staff makes these expenditures harder to determine.

Returning to the Kentucky example, their Request for Proposals called for the development of a totally performance-based assessment system for statewide use in selected grade levels and subject areas. The Kentucky Department of Education and external consultants estimated costs by assuming a collaborative development involving a small group of expert classroom teachers familiar with the target content, and measurement specialists sensitive to the desirable administrative and psychometric properties of performance tasks. Task 3 included the development of scrimmage events pretested in the state and therefore less secure, designed for use by schools in grade levels other than 4, 8, and 12. The estimated expenditure was $193,843 ($274,963) for 35 exercises,

or $5,500 ($7,801) per exercise. Task 4 included the development of secure exercises pretested outside of Kentucky and designed for statewide administration in Grades 4, 8, and 12. Task 4 represents a more likely scenario for states operating high- stakes assessments under an accountability mandate. The education department estimated Task 4 expenditures as $3,789,150 ($5,374,849) for 602 tasks over 5 years, or $6,294 ($8,928) per task.

However, a report found that first-year expenditures for developing these tasks exceeded initial estimates. Developers attributed unanticipated costs to a lack of existing examples for modification or to serve as models as well as significant additional time for developing of scoring rubrics. They also noted easier development of performance tasks for math and science than for social studies.

Hardy (1995) offers other examples of development estimates which confirm the per-task amount found in Kentucky. Educational Testing Service (ETS) estimated that the development of a 20-minute performance task for science (from the first draft to first administration to students in a pretest) required around 75 hours for an experienced test specialist and cost about $5,000 ($7,092) for development to this stage. This estimate only included an initial draft of scoring rubrics and not costs associated with refinement of rubrics as part of the scoring process. One middle-school mathematics project, was estimated to cost $6,410 ($9,092) per task, but these costs included a variety of instructional materials to complement the performance tasks. Using these examples as a basis, Hardy (1995) estimates that a basic unit of four performance tasks (1-hour testing time) in each of two subjects would have a unit cost between $45,000 ($63,831) and $60,000 ($85,109). Increases would occur from factors such as expected student outcomes not well-defined, a maximum involvement of teachers as developers, large tryout samples, and scaling or equating.

Hardy (1995) then provides estimates of the expenditures for administration of performance assessments. These include expenditures for any materials required to administer an assessment to students as well as for any special training for teachers, test coordinators, or other school personnel involved in the administration of assessment tasks. Test-administrator time is included as well.

In 1991, ETS developed four prototype assessments in elementary science for the state of Georgia. The most expensive task was an exercise requiring students to test and then identify six different mineral samples. The assessment kit included these samples, each labeled with a number, a small magnifier, a nail, a 2-inch square of glass, and a 2-inch square of ceramic tile; a single kit cost $9.00 ($14.28 in today's currency). The least expensive kits cost $0.70 ($1.11) for an exercise in designing a shipping carton to hold bars of soap and materials. The kit included a block of wood the size of a bar of soap and a 6-inch plastic ruler. The two other exercises developed for this project cost about $1 ($1.59) per kit and $4 ($6.35) per kit. Hardy (1995) also reported that NAEP science assessment kits were reported to range from $1.98 ($3.14) to $13.50 ($21.40) per

kit. In the case of materials, unit costs decrease as quantity increases. From these figures, Hardy estimated that the materials and distribution for a basic unit of four tasks at $3 ($4.76) to $5 ($7.94) per student, with increased manipulatives requiring higher dollar amounts.

Examining staff personnel costs, Hardy (1995) noted that NAEP and Kentucky sent specially trained task administrators to schools to test a sample of the enrolled students. While this approach saves local staff-training costs (which could have benefits if tied to PD activities), it adds to the expense of paid assessment administrators. Kentucky assessment staff administered one performance task to each of the 140,000 students at an average labor-and-travel cost of about $5 ($7.29) per student. Hardy (1995) estimates training costs at $150 ($218.80) per teacher, with personnel expenditures rising through the use of external examiners and the use of teachers as observers or raters of student performance.

In the third section of his analysis on scoring performance assessments, Hardy (1995) included training for teachers, other professionals and, in some cases, clerical staff to assign numerical scores, narrative comments, or other forms of evaluation to student responses to assessment tasks. Costs in this area are significant because of manual scoring. However, these costs require a new analysis under our new framework that provides a larger cost-benefit picture that applies to teacher-moderated scoring, among other pieces. Hardy (1995) offered a long list of performance-assessment scoring elements (Table 5) that informs estimates of a basic unit of narrative responses, holistically scored at a rate of 12 minutes per student with a basic unit cost of $3 ($4.38) to $6 ($8.75) per student.

The following factors affect scoring costs: multiple scoring and score resolution; the length and complexity of student response; analytic scoring and diagnostic reporting; scoring requiring special content expertise; and maximum involvement of classroom teachers (Hardy, 1995). Hardy (1995) also reports research which points out that as much as 60% of the costs of performance assessments apply directly to teachers for participation in scoring. This finding again mirrors our assertion about the real location for the cost of performance assessment—teacher PD and training—and the need to reframe the cost-benefit analysis.

Table 5. Scoring Estimates for Performance Assessments Cited by Hardy (1995)

Assessment	Scoring	Cost	Study
Connecticut Assessment of Educational Progress: 25- minute essay	Twice holistically (does not include staff costs for recruiting raters, procuring scoring sites, training table leaders, and selecting rangefinder papers and other categories)	$1.13 ($1.65) per student	Baron, 1984
Research study for SAT: 45-minute essay	Scored once holistically	$0.54 ($0.79) to $1.47 ($2.14) per student	Breland, Camp, Jones, Morris and Rock, 1987
California Assessment Program: 45-minute essay	Scored twice	$5.00 ($7.29) per student	Hymes, 1991
College Board English Composition: 20-minute essay	Scored twice	$5.88 ($8.58) per student	U.S. Congress Office of Technology Assessment, 1992
Geometry Proofs	Not reported	$3.00 ($4.38) per student	Stevenson, 1990
Kentucky Assessment: On-demand tasks in a variety of subject areas	Total scoring time per student: 12 minutes	$3.00 ($4.38) per student	Hill and Reidy, 1993

NECAP Cost Analyses

Just as historical estimates of the cost-of-performance assessment vary widely depending on accounting procedures and cost measurement, so do present estimates. The experiences of the states participating in the New England Common Assessment Program (NECAP) suggest a new route for offering a high-quality assessment at an affordable cost. Four states—New Hampshire, Rhode Island, Vermont, and (most recently) Maine—have banded together to form NECAP. All four participate in the consortium to provide their students with an assessment system including more than multiple-choice items.[3] By signing a joint contract with the testing firm, Measured Progress, to develop, administer, score, and report the results of their tests, the NECAP states are able to lower their individual costs, making it feasible for each to employ what might otherwise be a prohibitively expensive form of assessment. According to the terms of their contract with Measured Progress, each state must pay one fourth of the total fixed cost of the assessment and its portion of the variable cost based on how many students it tests.[4] As a result, each state pays only a fraction of what it might have

3. This explanation comes from phone interviews on November 23, 2009, with officials in the assessment offices for Vermont and New Hampshire (Michael Hock and Tim Kurtz, respectively).
4. Tim Kurtz, from New Hampshire, helpfully provided this explanation in our November 23, 2009, phone conversation.

paid in fixed costs were it to conduct the same testing program on its own. Because fixed costs comprise roughly 20%—$2.1 million—of the total $9.4 million NECAP price tag, this represents sizable savings for every state in the consortium.

In addition to dividing the fixed costs of the program, the NECAP states save money by realizing a number of economies of scale, as predicted by the GAO's 1993 cost study. In fact, New Hampshire's Assessment Director believes that economies of scale exist in the process of scoring open-ended items, which cost the state considerably more to score than multiple-choice items and represent one of the primary factors driving the somewhat higher price of performance assessment. According to New Hampshire's Assessment Director, "The first 1,000 constructed-response items are a lot more expensive to grade than the last 1,000" because, as graders become more experienced at scoring an item, they increase their efficiency and reliability.

The NECAP tests, while not solely performance-based, do include a substantial percentage of constructed-response questions in addition to multiple-choice questions. Item writers design these constructed-response items to elicit higher-level thinking and they account for about half of a student's score on the reading and math test. NECAP states administer the reading and math assessments during October in Grades 3-8 and 11.[5] As part of the same contract with Measured Progress, the states also administer a writing test to students in Grades 5, 8, and 11. In Grades 5 and 8, the writing assessment includes 10 multiple- choice questions, three constructed response items, three short-answer items, and an essay-length writing prompt. In Grade 11, the writing assessment includes one common, essay-length writing prompt and one equating or field-test writing prompt (Measured Progress, 2009). The total cost of developing, administering, scoring, and reporting these assessments for the 2009–2010 school year is roughly $29 per student tested and $12 per test administered.[6]

Beyond considerations of price, evidence from NECAP appears to illustrate the potential for improved teacher professional development that could result from switching to performance assessment, an important benefit notoriously hard to quantify. Rather than detracting from teaching time, the Director of Assessment for Vermont believes that NECAP testing has become "an embedded part of the curriculum," giving teachers valuable data to identify gaps in student knowledge. In fact, Vermont encourages its teachers to use released NECAP items as a model for crafting their own assessments, which state officials argue leads to the development of higher-quality classroom assessments capable of producing more meaningful results. Such improved teacher professional development and practice represents a potential benefit of switching to performance assessment, one which has the capacity to offset the marginal increase in price that states will likely incur by switching.

5. Maine does not participate in NECAP testing at Grade 11. Nor does Maine participate in the science portion of NECAP.
6. This estimate is based on a total of roughly 325,000 students tested and 780,600 tests administered.

Estimates Based on Assessment Company Revenues and State Budget Expenditures

Hoxby (2002) analyzes assessments in reference to accountability systems that include (in her definition) testing, standards for comparison of test results, report cards that relay this information, and information on schools' level and use of resources. Her study, done at the beginning of NCLB, includes figures on both multiple-choice and different types of performance assessment. However, it focuses on expenditures and does not provide the cost-benefit analysis outlined above. To analyze financial aspects of assessment, Hoxby (2002) examines two sources: revenue to assessment companies and state budget expenditures. Revenues to companies include sales of tests, standards-related materials like curriculum guides and criteria, and services associated with accountability such as consulting. State budget expenditures include payments to test-makers, cost of running an accountability office, salaries of accountability personnel, cost of publishing reports, cost of ongoing redevelopment and evaluation of the system, cost of consultants, and reimbursement to school districts for costs imposed on them.

Nationally, in 2000, total revenue to companies was $234.1 ($293.9) million, which was $4.96 ($6.23) per student or .07% of total spending on education. State variations occur based on four main differences: the size of the accountability office; states requesting tests that have curriculum guides specific to the state—additional costs still vary to the degree that a state desires an idiosyncratic test; the stage of test/assessment development (early on, states will need more personnel); and the size of the state (since fixed costs are shared by a smaller population in small states than large states). Comparisons between states, even those using strictly multiple-choice tests, lack precision without accounting for the listed differences. In this comparative analysis, Hoxby (2002) suggests that costs can be overstated because accountability systems have public support. This may provide states with incentives to exaggerate the share of their department of education's overhead associated with accountability.

Hoxby (2002) provides two state examples germane to this report: California and Kentucky. California's accountability system had a total cost of $19.93 ($25.02) per student. The system included the Standardized Testing and Reporting [(STAR) current system under modification], the Golden State Exams [new system being designed], and a high school exit exam. At the time of the study, California paid for the development of multiple tests. The STAR tested Grades 2–11 in reading, language, spelling, and math and Grades 9–11 in science and social studies. The Golden State Exams tested Grades 9–12 in reading, language, written composition, mathematics, science, Spanish, and history and social science. In addition to test development, the assessment system required seminars for school staff, experts to explain the system, experts to evaluate how the system is aligned with California's standards, an ongoing review of the system, and a few additional tests (English language development and career assessment).

Expenditures for personnel to administer the system included: new department of education staff for the Public School Accountability Act, at $0.31 ($0.39) per student and STAR tests at $0.07 ($0.09) per student, consultants for the Public School Accountability Act at $0.04 ($0.05) per student and high school exit exam at $0.02 ($0.03) per student, and test experts for STAR and high school exit exam at $0.06 ($0.07) per student. The salaries and fees of all these personnel spread over all the students in the state amounts to $0.50 ($0.63) per student. The costs of complementary activities were: website at $0.17 ($0.19) per student), test integrity at $0.03 ($0.04) per student, alignment with state standards $0.50 ($0.63) per student, reliability testing at $0.05 ($0.06) per student, test development including that of Golden State Exams at $1.98 ($2.49) + $0.25 ($0.31) per student, and assessment review $0.62 ($0.78) per student. We outline these expenditures in detail because they provide a finer-grained illustration of per-capita costs for some of personnel required in assessment, although California did not administer performance assessments at this time. It should also be pointed out that because of the large number of students in California—over six million—there are certainly economies of scale in many of the expenditure items for development and other activities that would not be found if smaller states developed testing programs on their own.

Kentucky, on the other hand, did have an assessment system that included both portfolio assessment and longitudinal assessment. These elements contain a high degree of individuation for each student and included features requiring intensive use of experts' time in developing and managing the process, as well as teacher time for scoring. The total cost per student in FY 2001 was $16.57 ($20.80). Having discussed Kentucky's system above, we present it here as a comparison to the California system. Overall, the FY 2001 cost was less in Kentucky for a performance-based assessment than in California for a largely multiple-choice testing system. Although California had some additional testing elements (for example, its exit examination), both states tested multiple subject areas at the full range of grade levels required by NCLB.

An Evidence-Based Model of School Finance

One model that offers some insight into the costs of performance assessment is the evidence-based model of school finance adequacy developed by Odden and Picus (2008). Their model lays out a research-based approach to the organization of schools that often changes how certificated staff are used and provides funds for instructional supplies and materials. Generally, the model enables a school to implement the 10 strategies outlined by Odden (2009) that have been identified as frequently leading to strong gains in student performance. Among the strategies most aligned with strong performance-assessment practice are: a focus on planning and collaboration time for teachers, large investments in professional development that include additional paid days for teachers to meet during the summer to plan instruction (and the measures of their success in instruction), funds for instructional coaches to help teachers analyze assessment data and

improve instruction, and money to purchase the contract services of experts as identified by the school or district.

The actual cost of an evidence-based system varies substantially from state to state depending on the current level of spending and the number of certificated personnel in each district and school. In some states there are nearly enough certificated staff to fill the roles identified in the model while, in other states, additional staff may be required. Moreover, the assessment aspects of the model are one part of a systemic view of improving schools (Odden & Archibald, 2009; Odden, 2009), making it hard to distinguish the costs of the assessment system by itself.

That said, there are some direct expenditures a school or district must make to implement any assessment program. In our work in a number of states, we have estimated this to be approximately $25 per student, hardly a major expenditure compared to current levels of per-pupil spending in the states (see for example, Odden et al., 2006; Odden et al., 2007; Picus, Odden, Aportela, Mangan, & Goetz, 2008). This expenditure would include resources for testing materials, and enough funds to purchase an online system such as the NWEA's MAP tests, which cost approximately $7 per student. It does not include the "costs" of staff (either new positions or replacement of alternative activities by staff) for the implementation of an assessment system.

Conclusions

Research on estimating the costs of performance assessments in the United States can help inform new systems of assessment, especially if we use a framework that can distinguish between expenditures and costs, and can incorporate the student and classroom benefits of various kinds of assessment systems.

The key to completely identifying the costs and benefits of assessment programs is to understand how personnel time is used in the development, design, preparation for, administration, and evaluation of assessments. The lion's share of costs is the personnel time devoted to these steps of the assessment; the benefits accrue to the extent that the assessments help educators support and improve student learning. Research on the total costs of assessments then needs to focus on how personnel time is reallocated for different assessment strategies, and the benefit measured by improvements in the quality of instruction and the outcomes of those assessments.

If one were simply to look at the expenditures devoted to various forms of performance assessment, one would find that the expenditures as a component of a school district's budget are quite low. Yet a comprehensive system of formative, benchmark, and summative assessments requires considerable time on the part of teachers, school site leadership, and a central office. While current standardized tests are often viewed as reducing time for learning, because they are remote proxies from what actual student work,

curriculum-embedded performance assessments that provide learning experiences are typically viewed by educators as enhancing instruction, rather than impeding it—and research suggests that this is often the case. While any assessment program has considerable costs in terms of the personnel time devoted to conducting, evaluating, reporting, and using assessment results, useful information—about what students know and can do—and support for teachers' understanding of standards, curriculum, teaching, and learning are important benefits of high-quality assessment programs.

References

Baron, J. B. (1984). Writing assessment in Connecticut: A holistic eye toward identification and an analytic eye toward instruction. *Educational Measurement: Issues and Practice, 3,* pp. 27, 28, 38.

Belfield, C. (2000). *Economic principles for education: Theory and evidence*. Cheltenham, UK: Edward Elgar.

Boudett, K. P., City, E. A., & Murnane, R. J. (Eds.). (2008). *Data wise*. Cambridge, MA: Harvard Education Press.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., &Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. I I). New York: College Entrance Examination Board.

DuFour, R., DuFour, R., Eaker, R., & Many, T. (2006). *Learning by doing: A handbook for professional learning communities at work*. Bloomington, IN: Solution Tree.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: RAND Corporation.

Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education, 8*(2), 121–134.

Hartman, W. T. (2002). *School district budgeting*. Washington, DC: Association of School Business Officials, International.

Hill, R., & Reidy, E. (1993). *The cost,factors: Can performance based assessment be a sound investment?* Manuscript submitted for publication.

Hoxby, C. (2002). *The cost of accountability*. Working Paper 8855: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w8855

Hymes, D. L. (1991). *The changing face oftesting and assessment* (Critical Issues Report Stock No. 021-00338). Arlington, VA: American Association of School Administrators.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis (2nd ed.).*. Thousand Oaks, CA: Sage Publications.

Measured Progress (2009). *New England Common Assessment Program: 2008-2009*. Technical Report. Dover, NH: Measured Progress.

Mishan, E. J., and Quah, E. (2007). *Cost benefit analysis (5th ed.)*. New York, NY: Routledge.

Monk, D. H. (1990). *Educational finance: An economic approach*. New York: McGraw Hill.

Monk, D. H. (1995). The costs of pupil performance assessment: A summary report. *Journal of Education Finance, 20*(4), 363–371.

Odden, A. R. (2009). *10 strategies for doubling student performance*. Thousand Oaks, CA: Corwin Press.

Odden, A. R., & Archibald, S. J. (2009). *Doubling student performance… and finding the resources to do it*. Thousand Oaks, CA: Corwin Press.

Odden, A. R., & Picus, L. O. (2008). *School finance: A policy perspective, (4th ed.).*. New York, NY: McGraw Hill.

Odden, A., Picus, L.O., & Goetz, M. (2006). *Recalibrating the Arkansas school funding structure*. Prepared for the Adequacy Study Oversight Sub-Committee of the House and Senate Interim Committees on Education of the Arkansas General Assembly. North Hollywood, CA: Lawrence O. Picus and Associates.

Odden, A., Picus, L. O., Archibald, S., Goetz, M., Mangan, M. T., and Aportela, A. (2007). *Moving from good to great in Wisconsin: Funding schools adequately and doubling student performance*. Madison: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education. Retrieved from http://www.wcer.wisc.edu/cpre/finance/WI%20March%201%202007%20Adequacy%20Report1.pdf

Picus, L. O. (1994). *A conceptual framework for analyzing the costs of alternative assessment* (Technical Report No. 384). Los Angeles, CA: Center for Research on Student Standards, Evaluation and Testing. Retrieved from http://www.cse.ucla.edu/products/summary.asp?report=384.

Picus, L.O., & Tralli, A. (1998). Alternative Assessment Programs: What Are the True Costs? An Analysis of the Total Costs of Assessment in Kentucky and Vermont. Los Angeles, CA: Center for Research on Student Standards, Evaluation and Testing.. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH441new.pdf

Picus, L. O., Tralli, A., & Tasheny, S. (1996). *Estimating the costs of student assessment in North Carolina and Kentucky: A state level analysis* (Technical Report No. 408). Los Angeles, CA: Center for Research on Student Standards, Evaluation and Testing.. Retrieved from http://www.cse.ucla.edu/products/summary.asp?report=408.

Picus, L. O., Odden, A., Aportela, A., Mangan, M. T., & Goetz, M. (2008). *Implementing school finance adequacy: School level resource use in Wyoming following adequacy-oriented finance reform.* Prepared for the Wyoming Joint Interim Legislative Education Committee. North Hollywood, CA: Lawrence O. Picus and Associates.

Snyder, T.D., & Dillow, S.A. (2010). *Digest of Education Statistics 2009* (NCES 2010-013). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Stecher, B. (1995). *The cost of performance assessment in science: The RAND perspective.* Paper presented at the 2006 National Council on Measurement in Education, San Francisco, CA.

Stevenson, Z., Averett, C., & Vickers, D. (1990, April). *The reliabiliry of using afocused-holisticscoring approach to measure student performance on a geometry prooj:* Paper presented at the meeting of the American Educational Research Association, Boston, MA.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (Report No. OTA-SET-519; pp. 216, 243, 210). Washington, DC: U.S. Government Printing Office.

U.S. General Accounting Office (1993). *Student extent and expenditures, with cost estimates for a national examination* (Report GAO/PEMD-93-8). Washington, DC: General Accounting Office.

U. S. General Accounting Office (2003). *Title I: Characteristics of tests will influence expenses; information sharing may help states realize efficiencies.* Washington, DC: GAO.

U.S. Government Accountability Office (2009). *No Child Left Behind Act: Enhancements in the Department of Education's review process could improve state academic assessments* (Report GAO-09-911). Washington, DC: Government Accountability Office.

Wylie, C., & Lyon, C. (2009). What schools and districts needs to know to support teachers' use of formative assessment. *Teachers College Record*. Retrieved from http://www.tcrecord.org/content.asp?contentid=15734.

Linda Darling-Hammond, Co-Director
Stanford University Charles E. Ducommun Professor of Education

Prudence Carter, Co-Director
Stanford University Associate Professor of Education and (by courtesy) Sociology

Carol Campbell, Executive Director

Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu

http://edpolicy.stanford.edu