C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Economic Analysis of Testing: Competency, Certification, and "Authentic" Assessments

CSE Technical Report 383

James S. Catterall and Lynn Winters
CRESST/University of California, Los Angeles

August 1994

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

This research was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

ECONOMIC ANALYSIS OF TESTING:
COMPETENCY, CERTIFICATION, AND "AUTHENTIC" ASSESSMENTS[1]

James S. Catterall[2] and Lynn Winters[3]
CRESST/University of California, Los Angeles

Abstract

This report explores the application of cost analysis to testing and assessment in elementary and secondary schools. A case for economic analysis of testing is first outlined. The case rests on the importance of test information, the magnitude of resources devoted to testing, and the relevance of an economics of information model to testing decisions. Second, the common modes of economic analysis attached to this model are discussed: cost-benefit and cost-effectiveness analyses. The presentation is carried out through examination of systemwide tests of pupil achievement and teacher competency. Finally, the contemporary press for more "authentic" or performance-based assessments is explored using the lenses of economic models.

Introduction

This report explores the application of cost analysis to testing and assessment in elementary and secondary schools. First, a case for economic analysis of testing is outlined. The case rests on three points—the importance of test information, the magnitude of resources devoted to testing, and obvious relevance of an economics of information model to testing decisions. Second, the common modes of economic analysis attached to this model are discussed: cost-benefit and cost-effectiveness analyses.

---

[1] Portions of this report contain revised and updated discussions from "Estimating the Costs and Benefits of Large Scale Assessments" (James S. Catterall, 1990*, Journal of Education Finance, 16*, 1-20).

[2] James S. Catterall is Associate Professor at the UCLA Graduate School of Education and Head of the Administration, Curriculum, and Teaching Studies Division.

[3] Lynn Winters is currently a Project Director at the Center for Research on Evaluation, Standards, and Student Testing (CRESST). This article was completed during her former tenure as Director of Research at the Galef Institute of Los Angeles.

The presentation is carried out through examination of systemwide tests of pupil achievement and teacher competency, which have grown in both popularity and controversy in recent years. Finally, the contemporary press for more "authentic" or performance-based assessments is explored using the lenses of economic models. The discussion focuses on the promises and limitations of economic analysis rather than any weight of accumulated evidence, in part because no such weight has registered. We conclude that economic reasoning has an important place in testing decisions, but that its role has failed to match the promise of theory. At the same time, today's lively debates over the course of testing and assessment in American schools embrace unmistakable cost-benefit themes and are likely to be the focus of clearer discussions of costs and effects as experience with these assessments accumulates.

The Brief for Economics and Testing

At first blush, the quest for instructionally-important or policy-relevant information through testing and assessment seems a uniquely qualified candidate for economic scrutiny. One reason is that the benefits of testing are alleged to be vitally linked to the delivery of appropriate educational services and to the motivation of learners and educators. The information elicited in tests has potential value to a number of audiences: students, parents, teachers, administrators, legislators, and society. Tests are supposed to help identify the needs of learners, assess whether learning has taken place, and appraise the effectiveness of schools or systems. A second reason is that our education systems make substantial contributions of human and material resources to various forms of testing. In a climate of fierce competition for scarce public and private resources, questions of return to investment in assessment seem very important.

This establishes a natural role for the economist, who probes relationships between costs and benefits, costs and effects, and costs and more general notions of utility in human endeavors. Economics also offers a subfield directly concerned with questions of returns to test-like activities, namely the economics of information. This paradigm recognizes the value of information and the consequences of decisions to pay for it. Like any good, information entails both return and cost. In applying microeconomic reasoning to information seeking,

economists can illuminate questions concerning what resources a decision maker should allocate to a search for information and what patterns of cost and benefit are associated with information collection. While this literature addresses consumer behavior, market information, and questions like how long to search for a lower price, the economics of information perspective has direct application to testing and assessment. By its very nature, an assessment is a mechanism for collecting useful information that exacts a toll on both system budgets and resources contributed indirectly.

Application to Large-Scale Assessments

The term *large-scale assessment* refers to attempts to judge the status of student learning or the value of key contributing resources, such as teachers, across an entire educational system. These testing programs, which include national and state student achievement assessments and teacher certification tests, can cost sponsors millions of dollars in direct costs as well as induce a variety of less visible costs. The focus of this section is on how the costs and benefits of large-scale assessments can be appraised and the usefulness of such cost analysis studies to policy makers and educators. Although these questions are not new to the education research community (see Solmon & Alkin, 1983), the presentation does bring recent data to bear on them. One source of insights is a study of the Texas teacher certification test (Shepard & Kreitzer, 1987a, 1987b). Another is a national study of pupil minimum competency testing (Catterall, 1989).

Cost-Benefit Analysis and Large-Scale Assessments

Cost-benefit analysis (CBA) refers to the comparison of the *costs* of an enterprise to the *benefits* of that enterprise *where both costs and benefits are measured in monetary terms*. Simply put, CBA requires the analyst to estimate dollar equivalents for both costs and benefits of an activity under scrutiny. This facilitates clean analytical comparisons such as ratios of benefits to costs, the amounts that benefits exceed or fall short of costs, or the implied rate of return on dollars invested in activities (Levin, 1983). There is distinct convenience in achieving dollar quantities for both costs and benefits of public programs. The user of CBA findings may conclude that an activity is worth doing for a very compelling reason: The benefits achieved exceed the costs borne to achieve

them. Or with such information one might, for example, be in a position to judge a project in public health more worthy than another in transportation on the grounds that the former produces a greater net return on the investment. These are potentially powerful comparisons and are fully justified when we reach consensus that the relevant costs and benefits have been identified and their values properly assigned.

When we try to apply CBA principles to large-scale assessments, however, we confront perplexing barriers to securing dollar values for many of their touted benefits. While useful monetary translations have been applied to selected effects of some assessments, it is unlikely that any dollar total will adequately characterize the full spectrum of effects resulting from a given assessment program. A listing of plausible effects of teacher certification tests (Figure 1) suggested by the work of Shepard and Kreitzer (1987a, 1987b) points to these difficulties.

The nature of the example effects of teacher certification tests—and surely there are more—precludes obtaining dollar values in straightforward or convincing ways. Thus, when we ask what we are getting in return for a certification test's costs, we are unlikely to find the economist's ideal answer. We cannot hope to say that the overall benefits represented as dollar values yielded by a competency exam or a teacher certification test exceed or fall short of the costs of administration. However, we can hope to approach this ideal in a way that may be relevant to the analysis of assessment policies. This is the subject of the next section.

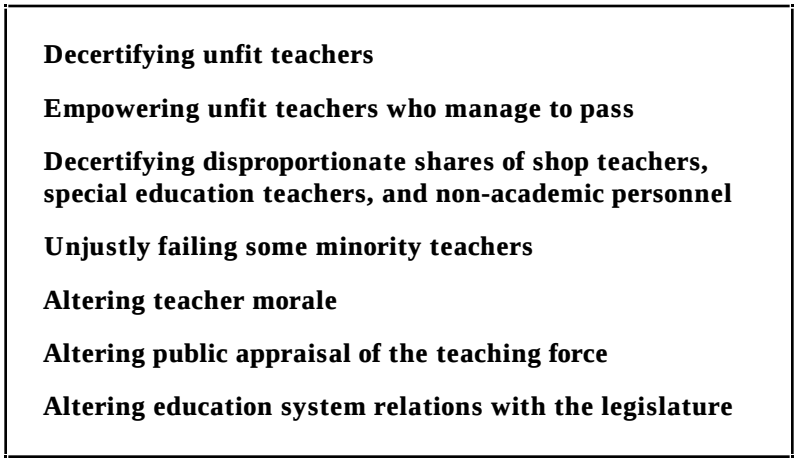

Figure 1. **Example effects of teacher certification tests.**

Partial cost-benefit analyses. Because of the non-monetary nature of many benefits, we argue against reliance on cost-benefit analysis techniques to appraise the overall worth of large-scale assessments. However, because some of the effects of these assessments can be thought to yield benefits with recognizable monetary values, a partial analysis using CBA techniques may be useful. That is, particular effects may be quantified in dollar equivalents and examined in relation to assessment costs. Under circumstances where these effects are deemed important or to involve large sums, the results can be revealing and decision-relevant.

One example of this is reported in Shepard and Kreitzer's (1987a) analysis of the recently administered Texas teacher test, the TECAT. This test, administered to more than 200,000 practicing teachers, was designed with a simple set of objectives: to ensure that teachers met minimum levels of literacy, to decertify teachers who could not meet these standards, and to convince Texans that their children's classrooms were staffed by capable teachers. The authors found that a host of effects, both planned and unplanned, accompanied the administration of the test. To some of these they applied cost-benefit analysis reasoning.

One such instance was comparing the costs of the assessment to the recaptured salaries of ousted teachers. The total public cost of the TECAT was nearly $36 million, amounting to nearly $30,000 per failed teacher. One "benefit" of removing these teachers was that their salaries totaling $25 million annually, an amount assumed to be wasted, would no longer have to be paid. The authors might have estimated these salaries over the remaining careers of these teachers for the analysis, a perspective that would probably have shown present value savings in the hundreds of millions of dollars. This perspective would have suggested a very high ratio of benefits to costs on this one dimension alone.

Another CBA example from this study was the authors' inquiry into the opportunity costs of spending some $30 million on the Texas test. Opportunity cost refers to the question "What purposes might have been achieved with an expended sum if the resources had been allocated in a different way?" That is, given their inherent scarcity, an important cost of assigning resources in one manner is the sacrificed opportunity to pursue the next best alternative. In assessing the TECAT, the authors offer the possibility that student learning

across the state might have benefited significantly from tutoring services that the TECAT'S $30 million price tag could have purchased—about 14 hours from every state teacher (both fit and unfit, of course).

Yet another example of a partial cost-benefit analysis is found in the analysis of the Shepard and Kreitzer study by Solmon and Fagnano (1990) which suggests that unfit teachers promulgate undereducation of pupils. This result leads to lower productivity and to social costs that can be estimated in dollar values (see Catterall, 1987). And an undeveloped but analogous line of reasoning applies to our recent research on minimum competency tests; here we found strong indications that tests required for the high school diploma may induce dropout responses among test failers (Catterall, 1989). These dropout tendencies could be translated to private and social costs (Catterall, 1987), and these sums in turn could be compared to the costs of the competency testing system. In each of these examples, neither the original authors nor we in our re-analyses claim to provide monetary estimates of the total benefits deriving from the tests under scrutiny. Rather, a selected focus is chosen that produces costs and benefits in dollar terms related to that analytic frame. This evidence may then play a contributing role in policy discussions concerning the wisdom of such assessments, an implicit goal of this sort of research. And where a benefit or cost is shown to be preemptive, a partial analysis may provide a very compelling argument for continuing or abolishing a large-scale assessment.

Cost-Effectiveness Analysis and Large-Scale Assessments

As illustrated in the discussion thus far, limitations in assigning monetary values to benefits dull the application of cost-benefit analysis to large-scale assessments. This problem does not haunt cost-effectiveness analysis. Unlike CBA, cost-effectiveness analysis (CEA) entails estimating program effects in their naturally-occurring units and then relating these effects to costs (Levin 1983). An example finding of a cost-effectiveness approach was contained in the TECAT study: that the test costing $36 million decertified 1199 teachers; one cost per unit of effect would read $29,703 per teacher expelled ($36 million/1199).

Cost-effectiveness analysis requires that effects be quantified, but not that these quantities be expressed in dollars. Relaxing the need to provide monetary estimates can bring formerly problematic outcomes into the analyst's purview. Note that under CEA, each of the plausible effects of teacher certification testing

listed above lends itself to measurement. The number of teachers decertified is a simple (or not so simple) count, the tendency to decertify minority teachers could be described by measures of distribution, teacher morale can be judged using scales constructed from interview responses, public confidence in teachers can be surveyed, educator relations with legislators could be assessed by examining state budget allocations or through more direct measures, and so on.

While comparing effects expressed in natural units to dollar costs avoids the limitations described for CBA, the tendency of large-scale assessments to have multiple effects limits the utility of CEA to decision makers. When one judges the effects of an assessment program in a half dozen or more important domains, the effects side resembles a shopping basket: W numbers of teachers bumped, X percentage decrease in teacher morale, Y units increase in reading achievement, Z percent of special education teachers decertified, and so on. Assuming that all of these effects can be measured satisfactorily, this basket characterizes what the assessors gained for their money. But this information does not necessarily help with decisions to continue or modify the assessment practice.

One obvious shortcoming derives from *not* securing dollar values for the effects, a problem set up in choosing CEA over CBA in the first place. Effects estimates in their natural units do not provide a ready answer to the question of whether the assessment was worth it. We may be able to say that we gained 2 apples and 3 oranges and lost 1 plum, but whether this is worth the dollar spent must be resolved by human judgment.

A second shortcoming of CEA is the difficulty of comparing baskets of anticipated effects where competing assessment strategies are contemplated. The importance of such a decision context cannot be overstated, because choosing alternative means to a desired set of ends lies at the heart of public policy making, including deciding on large-scale assessment policies. It is likely that the baskets of effects associated with competing assessment policies, such as two differing schemes for testing teachers, will have dissimilar and ambiguously valued contents. Only if one basket has more (or the same amount) of the desired effects, and less (or the same amount) of the disliked effects, is a preference for one of them clear. And even this case assumes agreement on which are liked and which are disliked among effects. For example, CEA does not provide any assistance with the problem of comparing the decertification of 100 incompetent teachers on the one hand with a 0.13 standard deviation

reduction in teaching force morale on the other. In the idealized world of CBA, where dollars are assigned to all effects, we could make such a comparison.

Thus it appears that while cost-effectiveness analysis methods allow us to transcend an inability to construct comprehensive dollars-only cost-benefit analyses, CEA techniques have their own complications where the provision of policy-making advice on large-scale assessments is concerned. Yet both CEA and CBA perspectives have been shown to offer potentially useful information to the policy maker. We turn now to an observation suggesting one of the reasons why we have not seen more use of these techniques in research and analysis on large-scale assessments—the difficulty of determining some of the effects in the first place.

Assessing Program Effects

Attributing general educational or social effects to small contributors, such as testing, within large complex systems is a daunting analytic prospect. Such is the task implied when we list among the outcomes of large-scale assessments such effects as pupil learning, teacher morale, public confidence in education, and education system relations with the legislature. When we hope to quantify such global effects, either in naturally occurring units (for CEA) or in dollar equivalents (for CBA), the first order of business is the inherent research or evaluation problem. These outcomes can only be truly accounted for in the context of the arrays of forces that play on each of them. Research that can convincingly partial-out the effects of a mere educational assessment on any of these outcomes would have to be clever, powerful, and probably very expensive.

One approach is the search for proximate effects undertaken by Shepard and Kreitzer in the study noted above. They employed a longitudinal design to assess the effects of the TECAT. They examined conditions before and after the test, largely through perceptions of participants as revealed in interviews. This design may be the best that can be achieved without massive resources. It has the advantages of tapping into stakeholders for whom effects are both important and tangible and who are in a position to help isolate the effects attributable to the testing program. It has disadvantages in the form of possible response bias introduced because respondents may surreptitiously favor or disfavor the assessment system for undisclosed reasons.

A quasi-experimental design is potentially applicable to our effects assessment questions. This would be to analyze differences across educational systems to judge the influence of a large-scale assessment effort. One such possibility is comparing a state with a teacher certification test to one without; another is pooling states for multivariate analysis. Unfortunately, the likelihood that the assessment system would be the only difference between two large settings (among factors influencing the outcomes in question) seems very small. This limitation hampers two-state designs. For a multivariate attempt, it may be impossible to identify sufficient cases to sustain an analysis. By large-scale we have generally meant state-level assessments, and this limits the population to 51. This would allow for a very constricted set of determinants of any global outcomes in question, and the resulting models would probably be very poorly specified.

These perspectives do not engender optimism. Readers may be justified in a retreat to the more selective, partial analyses discussed above. Analysis that is restricted to the immediate numbers generated when assessments are conducted, such as teacher firings or diplomas denied, may still be useful to decision makers. The most credible and attainable data appear to focus on proximate effects that are easy to detect and which also represent logical consequences of an assessment.

Paucity of CBA and CEA Studies

Educators, policy makers, and citizens alike believe that tests are to varying degrees capable of measuring the knowledge and skills held by individuals. This in turn legitimates the use of tests to assign individuals, such as students and teachers, to particular knowledge-based or skill-based statuses. A 10th-grade standing, a diploma, a teaching post, or a pink slip come to mind. And if high stakes attach to the assessment, the reasoning is that testing will induce desired behaviors in those to be tested. Consequences widely expected from tests that determine status include studying harder, attending and being more attentive in school, and correcting anticipated or revealed deficits.

It follows that policy makers and citizens can easily assume that tests will reap certain benefits. The tests used for large-scale assessments of both students and teachers are no exception. Teacher certification tests will cause teachers to brush up on their skills; and because these tests can be designed to

assess critical basic skills, they will point to those who are deficient in such domains. Tests required of pupils for high school graduation will cause students to tend to business; because these tests can be designed to assess the knowledge and skills needed by young adults, they will identify those who need to learn more before being granted a diploma. If these positive assumptions about large-scale student and teacher assessments characterize the political environment of these tests, the size and stability of the testing enterprise is a logical expectation. These assumptions also appear to suppress investigations that might pose challenges to them.

That unsubstantiated assumptions of benefits are actually made by testing policy makers and educational leaders is evident in our research on pupil competency testing (Catterall, 1989). This research focused on the minimum competency tests required for graduation from high school, a practice in more than half the states. The most graphic testimony is the dismal state of policy-relevant information concerning pupil performance on these tests. In our national study, we found not a single school, district, or state that tracks the subsequent performances of youngsters who fail part or all of a required graduation test, usually first taken in the 9th or 10th grade.

If a test is designed to yield positive outcomes for those who fail, such as well-targeted remediation and eventual passing performance, knowing whether this occurs and for whom might be useful for making decisions about the testing program. What is documented in most settings is limited to first-time pass rates at best. Where pass rates on readministrations are monitored, these statistics have scant meaning or utility because many students do not show up for retesting; an unknown but probably substantial number have dropped out of school.

Discrepancies between educator expressions and student data in our research on competency testing also support the assertion that the assumption of benefits tends to suppress their formal verification. For example, test coordinators, principals, and counselors expressed a common belief that the tests serve to motivate pupils. To examine the object of this belief, we asked students about their school's graduation testing requirements. We found that fewer than half of the 736 students in the 8 high schools (in four states) studied knew that passing a competency test was required for their graduation (see Catterall,

1990). About 45% of 9th graders knew of this requirement, a figure that grew only to 58% for 11th graders.

These perhaps puzzling levels of student awareness stand in stark contrast to the actual policies of the schools we studied, all of which required such tests and administered them for the first time in the 9th or 10th grade. More than 400 of the 736 students in our sample had in fact already taken the test, but many did not realize this. Our analysis of this awareness gap points to what we labeled a "testing blur" in American high schools—tests of varying descriptions come and go in the lives of students without ascribed meaning. Their sheer numbers may leave students unable to recall the nature or importance of any particular test. And without meaning, can tests motivate students?

Another touted benefit of pupil competency tests is their role in the remediation of student skills. In our study, educators universally heralded the capacity of competency tests to spot individual learning difficulties and point to corrective measures. But students were not so sanguine on this topic. In response to a question of whether competency test failers subsequently receive sufficient remedial help, only 59% of students in general and 54% of test-failers themselves answered yes. Students with low grades, those most likely to reap the benefits of corrective measures, were actually less likely to offer an affirmative response to this remediation question than students with high grades.

In a larger sense, our research indicates that educator and policy-maker assertions regarding the educational benefits of competency tests do not seem to rest on confirmatory data. Our findings also suggest that detailed examinations of actual benefits of these large-scale pupil assessments might present challenges to some of these assertions. The research does suggest that "benefits" or "effects" of educational interventions such as testing arise from the assumptions people hold prior to the intervention rather than from observed results.

Benefits and the Use of Test Information

Some benefits of testing are highly dependent on how testing information is used. On the surface, it appears that most large-scale assessments of a given type, such as pupil competency tests or teacher certification tests, are rather similar. They are oriented to basic skills, they use a pencil-and-paper, forced-

choice response format, and they certify or decertify using a criterion-referenced standard. But the benefits of similar tests conducted in similar ways in two different settings can vary tremendously—the variation comes with how the assessment information is used by educators and the education system.

Our research on the graduation test provided support for such a "benefit follows use" thesis. Consider, as in the previous section, whether the benefit of pupil motivation accompanies the administration of a required graduation test. One school we examined would post conspicuously in a central hall the names of competency test failers—an exercise in "humiliation breeds competence" by all appearances. Another school in our sample first tested pupils in 9th grade, a common enough practice, and then held off readministration of the test until 12th grade. This long-delayed retest seemed a curious exception to common practice. The purpose of this time gap was, in the words of the school principal, "to save the trouble of retesting transients and the many others who would simply leave school anyway" (Catterall, 1989). In sharp contrast, other schools reported establishing special remedial classes or other interventions to guide test failers to eventual success. The motivational (and educational) benefits of the competency testing programs across these schools would appear to be highly disparate.

The benefits of competency tests could vary for other reasons associated with the use of test information. Remedial classes spawned by competency testing may, on the one hand, have the effect of filling knowledge gaps and repairing skill deficits. On the other hand, such efforts may advance test-taking skills or simply involve teaching directly to known test items. The former sound like desired learning effects; the latter do not. Such alternative practices appear to herald specific and differing benefits across competency testing systems.

On the Costs of Large-Scale Assessments

The advertised cost of large-scale assessments is usually limited to the appropriation accompanying their adoption by the legislature. In the case of the TECAT discussed above, the reported appropriation was $4.8 million. By the time researchers Shepard and Kreitzer (1987b) finished identifying resource contributions to the TECAT, the public costs approached $36 million, and the total costs were about $78 million when induced private costs were included. What accounts for such dramatic differences?

These discrepancies occur because the budget allocations to develop and administer a large-scale testing program typically fall far short of the costs of various other ingredients required to make a system work. The TECAT analysis showed public costs of $26 million for the in-service day used by teachers to take the test and another $3 million in district-paid workshops. These costs must be added to the initial $4.8 million public appropriation which paid for full-time administrators, direct development costs, materials, and scoring costs. Induced private costs such as teacher study time and privately paid workshops are appropriate candidates for inclusion in a complete cost accounting.

A similar but smaller-scale example of the discrepancy between published and true resource costs of testing appeared in our study of a school district's pupil information system. In this case, the budgeted costs for a curriculum-matched testing system in the study district's elementary schools amounted to 80 cents per pupil, but the costs of all ingredients identified as necessary to its operation totaled $34.00 per pupil (Catterall, 1984).

The primary reason that this "ingredients" perspective is appropriate to a cost appraisal of large-scale assessments is that these ingredients represent the opportunity costs of being in the assessment business. If there were no large-scale assessment, the various resources identified, public and private, might be put to some alternative use. This is what is sacrificed where an assessment program is maintained, and this is what represents its true costs.

An additional reason that the true costs of large-scale assessments may exceed their published budgets is the likelihood that negative effects will accompany large-scale assessments and that these effects will induce real costs. Analogous to the suggestion of Solmon and Fagnano (1990) that teacher recertification can induce dollar benefits when the removal of incompetent teachers is tied to future student productivity, our research on competency tests reveals an example of a cost-inducing possibility. We found that failing a competency test shows a strong tendency to depress students' self-expressed chances of finishing high school (Catterall, 1989). If competency tests tend to push out youngsters who might have persisted and otherwise benefited from school, some of the costs of dropping out could be pinned on such tests. This accusation seems particularly justified in the cases cited above, where the test seems to be used in a degrading fashion (such as publicly identifying students who fail) or where students who fail do not receive corrective attention.

We have not advanced the estimates in our competency testing study to the point where we can attribute particular numbers or percentages of dropouts to competency test failure. But we are aware from previous research that a single high school dropout sacrifices more than $200,000 in lifetime earnings and induces social costs in terms of both lost tax collections and also higher needs for a variety of public services (Catterall, 1987). The regression coefficient for test-failure in our dropout-likelihood model could be translated (on the basis of other research using this construct) to expected increases in dropouts. This in turn could be "costed" according to the procedures used in our cost-of-dropouts analysis.

Summing-Up

The objectives of cost-benefit or cost-effectiveness analyses of large-scale assessments are the attainment of optimal economic or educational outcomes. In the dollar-driven case of cost-benefit analysis, the optimum is economic. Here we are particularly concerned with maximizing dollar returns on expenditures or with maximizing net returns. In the effects-driven case of cost-effectiveness analysis, the relevant optimum is most appropriately composed of educational values. For CEA, we occupy the analysis with the educational effects sought through assessment, and we become concerned with maximizing educational outcomes rather than dollar returns.

In both CBA and CEA, however, the analyses have similar implicit purposes: showing policy makers just what it costs to achieve a particular set of objectives, and demonstrating what planned or unplanned ends are served by their policies. These analytical models also aim at what choices might be made either to reach given goals with lower costs, or to attain more results for a given budget allocation.

The shortage of relevant studies is noteworthy. Only trace quantities of cost-benefit research regarding the tests we have discussed have been reported. This paucity of analyses synthesizing costs and effects is mirrored by the scarcity of reported studies on either costs or effects alone. Information on the opportunity costs of large-scale assessments, a commodity highly appropriate for decision making, is practically non-existent. And our knowledge of the effects or benefits of large-scale assessments can only be described as thin. Decision makers and educators seem rather content with an assumption-based appraisal

of large-scale assessment effects, or with abject uncertainty about effects. These are certainly inexpensive and undemanding sorts of information, at least in the short run. This circumstance also implies that other types of information must drive assessment policy decisions.

There do not seem to be insurmountable technical impediments to making material improvements on this state of information. Measurement of some of the suspected benefits may be difficult and require approximations; a consensus on what benefits to examine may need to be forged; and quasi-experimental research designs (comparative across systems or longitudinal within a system) may need to be developed. But even in a world where we acknowledge that the ideal CBA and CEA models cannot be satisfied, we can know far more about the relationships between costs and effects of large-scale assessments than we do at present.

"Authentic" or Performance Assessment

Our discussion thus far has viewed well-established minimum competency and teacher certification tests from cost-benefit and cost-effectiveness analysis perspectives. We turn now to a very recent movement in educational assessment which entails increasing demands to include more genuine performances in the tasks required on tests of student learning. Both the evolution of these pressures and the changes sought in testing practices are candidates for examination through the lenses of the economist.

Setting the Stage for Performance Assessments

Competency testing practices underlying the examples discussed above, as well as the practice of large-scale achievement testing more generally, have followed an evolutionary path attuned to economic efficiency. To begin with, long-standing behaviorist theories of learning and its measurement bolstered practices of simple and efficient assessment. These perspectives hold that learning occurs through mastery of a hierarchy of discrete skills in a sequential and linear fashion. Assessment in turn focused on the acquisition of these bits of knowledge which could easily be represented in brief multiple-choice test items. (Behaviorists of course also saw benefits in the rewards and sanctions produced by test scores.) Efficiency in testing was achieved as well though the practice of sampling from knowledge domains tested. If the realm of sixth-grade

mathematics contains 700 facts and skills, a random selection of 70 or so skills (stratified by level) included in a test would support defensible inferences about the degree to which all 700 were known. Even greater efficiencies were elicited in large-scale assessments where individual students in a school class received even smaller but differing subsets of test items to produce a portrait of achievement generalizable only to the class. This latter practice characterizes the California Assessment Program (CAP) tests which produce school-level test scores for Grades 3, 6, 8, and 11.

These selected response tests realized a variety of economies: A large knowledge domain could be represented by a relatively small number of items. A large number of people could be tested simultaneously and in a relatively short period of time. Inferences could be made about people based on a small number of "observations" (test scores). With the advent of optical scanning and computerized scoring and reporting, a large number of tests could be scored and reports generated in minutes rather than hours or days. And because of standardized test administration conditions and objective scoring procedures (i.e., human judgment was subsumed into a scoring key) high test reliability became the norm.

This reliability and the apparent "objectivity" of selected response tests contributed in no small part to the increasing acceptance of standardized tests for a wide variety of testing purposes, including the minimum competency and teacher certification tests described above as well as classroom assessment, school-level evaluation, and career and psychological appraisals.

And as the customer base for standardized selected response tests broadened, large test publishing companies (such as the Educational Testing Service and CTB-McGraw Hill) were born, which made it possible for economies of scale to be realized in test development, scoring, and reporting. For example, a school can now purchase a 12-test multiple-choice achievement test battery for about $6.00 per pupil and have the tests scored for an additional $1.00. With the addition of reusable test booklets, the annual cost approximates the marginal cost of test scoring. Institutions realize great economies in buying off-the-shelf tests and using them over long periods of time.

This history is characterized by increasing uses of testing—reporting scores for individual students, school classes, schools, school districts, and even states,

that is, an expansion of benefits to various audiences. The history also shows very low costs for large-scale multiple-choice tests in the form of cheap materials and scoring, and incidental administration time.

Creeping Disutility of Classical Testing

As selected response tests came into wide use for a multitude of purposes and decisions, the benefits realized in their use suffered an unmistakable pattern of erosion. The main reason seems to have been that the tests increased in importance and spawned a range of behaviors that tended to corrupt their purposes (Haladyna, Nolen, & Haas, 1991; Shepard, 1988).

An example may be seen in the California Assessment Program (CAP). At the outset, this statewide pupil achievement test seemed like a good bargain. The assessment cost the state approximately $2.00 per pupil for ongoing development and administration; at the school-level, the only significant resource commanded was the 35 minutes of potential instructional time devoted to the test. Schools in exchange received grade-level profiles of strengths and weaknesses in reading and math, information they could use to design instructional improvements. CAP served as a correction/feedback mechanism for individual schools.

But beginning in the mid 1970s, the climate for educational testing including CAP began to change. Perceptions that schools were failing rose; SAT scores persisted in a decline begun in the mid 1960s. Students were entering school unprepared or not fluent in English. Students unable to read were graduating from the public schools and returning with lawsuits. We were a "Nation at Risk." Schools became the focus of public attention and pleas for improvement, a situation continuing to the present. California Assessment Program scores began to appear in local newspapers and to be cited by Realtors as selling points. Finally, the test was used to "motivate" the schools. Schools received additional state budget allocations for test score improvements.

This evolution of a relatively innocuous rite of spring to a high-stakes event has brought unintended dis-benefits. Students cheated more as tests became important. Teachers or administrators altered answer sheets. The curricula narrowed to focus on test objectives. Test-taking skill-building activities and actual test preparation began to usurp instructional time; educators began to

accept the inevitability of test coaching. Teachers facing certification tests spent money for test reviews, spawning new industries.

The cost side of the ledger also began to swell. First, accountability pressures of the last decade have taken a toll on teacher and student time. Studies in the early 1980s indicated that high school students were spending 3-5 hours per week taking tests of all sorts (Burry, Catterall, Choppin, & Dorr-Bremme, 1982). More recent studies have documented instances in which teachers allocate weeks of instructional time to preparation for large-scale achievement assessments (Shepard & Dougherty, 1991).

Interestingly, during this period of increasing test use new conceptions of learning were emerging from cognitive psychology. In contrast to the behaviorist views of learning as the accretion of discrete bits of knowledge, cognitive psychologists proposed that individuals learn through active construction of knowledge. As thinking and learning processes began to hold a central position in cognitive conceptions of learning and as more and more educators shared such views, processes became an eligible target of testing as well.

Cognitive notions of learning were first adopted by the subject matter specialists and communicated to teachers through national curriculum reform movements; process-based writing, meaning-based mathematics, and "hands-on" science are well-known examples. Subject matter specialists married their experience with test-driven instruction to their cognitive views of what ought to be tested and how. The result was a now-widespread demand for "authentic" or performance assessment, meaning testing the actual performance of students in "real" settings rather than on sheets of paper with bubbles for pencil marks.

Performance assessment tasks differ markedly from the selection of preferred answers to the traditional questions contained in large-scale assessments. Performance assessment tasks range across such demonstrations as essays, oral presentations, projects, exhibitions, "think aloud" descriptions of problem-solving behaviors, and group processes such as dances, dramas, and murals.

Benefits, Costs, and Performance Assessments

While there are no published cost-benefit or cost-effectiveness studies of performance assessments, the debate has been carried out along the rough lines

of such analyses. In simple terms, performance assessments are thought to bring about benefits in the form of reducing the dis-utilities resulting from traditional testing practices described above, and the conferral of new types of instructionally-important and policy-relevant information. At the same time, the costs of performance assessments are thought to far exceed the costs of the traditional tests. So the movement seems to raise important questions of return to added investment.

Among the more important dis-utilities of testing thought to be reduced by performance assessment is the encroachment of testing and test preparation time into instructional time. Although proponents of the current testing system would argue that teaching to the test when the test is worthy is indeed good instruction, teachers and many educators make clear distinctions between time used for instruction (working on the local or state-adopted curriculum) and testing (preparing for and administering state or nationally mandated tests). Because they look much like instructional activities, performance tests in contrast are viewed as extensions of, rather than intrusions into, instruction.

The benefits expected of performance tests encompass those of selected response tests, but they extend to many new domains including curriculum renewal (gearing the curriculum to student capacity to perform rather than to test performance), increased student content acquisition (reversing the narrowing effects of teaching to small samples of tested content), increased quality of student thinking (asking students to solve problems rather than select best answers), and better guidance for staff development (securing information relevant to the improvement of teaching).

In exchange for hoped-for benefits, the costs of designing, administering, scoring, and recording performance assessments would far exceed the costs of traditional tests as we have described them. The ingredients list is long: staff training, development of task specifications and prompts, administration including observing, keeping running records, and compiling portfolios, and scoring. In addition, the record of performances such as a large painting or sculpture may be bulky and entail storage costs; other demonstrations are ephemeral, disappearing once the students have danced, delivered a speech, or performed a drama. Recording these on video is a conceivable response, but one requiring expensive and fragile equipment and considerable amounts of real time to review.

The countervailing view is that performance assessment not only does not increase the cost of testing but it leverages testing dollars by improving curriculum and instruction. This viewpoint is most clearly stated in a recent assessment RFP issued by the California State Department of Education:

> **Testing costs of alternative assessments, especially the staff development component, should be considered as a part of *curriculum costs* [our emphasis]. Teachers' renewed motivation and commitment to the Curriculum Frameworks should be viewed as a major element in the cost-benefit analysis. (Page 10, Request for Applications for the Alternative Assessment Pilot Project, AB 10, Quackenbush, 1990)**

Under this concept of costs, rater training and scoring costs would be designated as staff development, a line item in the curriculum budget at the state level and part of the operating expenses of the schools at the local level. The costs of performance assessments would be expected to increase when some of the traditional concerns of assessment are introduced. If performances are to be general representations of learning, multiple performances must be elicited since any one task is not likely to represent a generous sample of a learning domain. If judgment is involved in appraising performance, multiple judges must be provided.

Where have we come? This discussion has described contemporary pressures to render student assessments more "authentic." In general, these performance assessments appear to offer great opportunities to learn more about what students can really do, including things we hope students will be able to do well as adults. Advocates see prospects for benefits of varying type. It is also evident that typically conceived performance assessments would be much more expensive than traditional tests.

Just how much of what sorts of benefits and costs might accrue to performance assessments would depend, of course, on the design of specific performance tests and testing systems. The field has begun experiments with various forms of performance assessment, and we are beginning to see in narrative form the building blocks of cost-effectiveness analyses. One block, of course, is the cost of scoring. At the state level, the most commonly used strategy for scoring is to use some sort of sampling of students, with the generation of school or system scores in mind. Vermont instituted a mathematics portfolio assessment that used a two-tiered sampling system to

minimize scoring costs. A random sample of student portfolios in a region were read in a local scoring center; the state assessment consisted of a random sample from the local centers. In another example from trials for the California state assessment, each student's 45-minute essay does get read, but most get only one rating. Rater training takes between 2 and 4 hours and individual papers require between 2 and 5 minutes to read. In the best case, where papers are relatively brief (students write differing amounts in response to a test prompt), average costs can be held to $3.00 to $5.00 per paper.

If, however, performance assessments are used to support high-stakes decisions in the way of current large-scale assessments, scoring costs could increase dramatically. For an important decision such as non-graduation to be made, an essay would need at least two or three ratings and more than one essay per student would be required, certainly if student essay performance were the decision criterion.

We are also learning about the costs of performance assessments through experimentation. For example, experiments with hands-on science assessments are beginning to produce information on the outcomes and ingredients of administering both hands-on and computer-simulated performance tasks in science learning, which could be costed (Shavelson, Baxter, & Pine, 1990). These developments suggest that we are likely to see more attempts to draw these new assessments into cost-effectiveness types of analyses in the coming years.

Where are we headed? Our discussion of economic analyses of performance assessment suggests that there are lessons we have learned from the past that will be valuable in appraising the benefits of these measurement techniques; and there are lessons still to be learned.

1. The ingredients approach to cost accounting applies in the case of performance assessment, but there is little agreement about whether the old categories of testing costs for development, administration, rater training, and scoring should all be included as testing costs for these new assessments. The educator point of view tends to blur the distinction between assessment and instruction in this environment.

2. Some of the dis-utilities associated with contemporary large-scale assessments are engendered by their high stakes. With no high stakes performance tests yet in widespread use, it is not clear whether the purported

reduction of these disutilities would actually accrue to performance tests. Perhaps they accrue to the present context of their use.

3. Both the difficulties of and promise in conducting cost-benefit or cost-effectiveness analysis appear to hold for performance assessment as well as large-scale assessment. For cost-benefit analysis, however, performance tests have a potentially more inclusive basket of returns—more candidates for partial cost-benefit analyses or analyses of effects bundles in their naturally measured units.

4. Performance assessment, with its clearer sense of returns to the classroom teacher, may serve to help educators understand the utility of economics of information perspectives as the debate proceeds. That these new assessments will be relatively costly is apparent to all. That they have returns over and above current tests is presently assumed. Establishing the linkages between the costs and benefits may be an important factor in the course of testing reform in the 1990s.

References

Burry, J., Catterall, J., Choppin, B., & Dorr-Bremme, D. (1982). ***Testing in the nation's schools and districts: How much? What kinds? To what ends? At what costs?*** **(CSE Tech. Rep. No. 194). Los Angeles: University of California, Center for the Study of Evaluation.**

Catterall, J. S. (1983). A theoretical model for examining the costs of testing. In L. C. Solmon & M. C. Alkin (Eds.), ***The costs of evaluation*** **(pp. 45-51). Beverly Hills, CA: Sage Publications.**

Catterall, J. S. (1984, April). ***The costs of instructional information systems: Results from two study districts.*** **Paper presented at the annual meeting of the American Educational Research Association, New York.**

Catterall, J. S. (1987). On the social costs of dropping out of school. ***The High School Journal, 71*****(1), 19-30.**

Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for high school graduation. ***American Journal of Education, 98*****(1), 1-34.**

Catterall, J. S. (1990) ***A reform cooled-out. Competency tests required for high school graduation*** **(CSE Tech. Rep. No. 320). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing. Eric Document No. ED 338 675**

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. ***Educational Researcher, 20*****(5), 2-7.**

Levin, H. M. (1983). ***Cost effectiveness: A primer*****. Beverly Hills, CA: Sage Publications.**

Shavelson, R. J., Baxter, G. P., & Pine, J. (1990, October). ***What alternative assessments look like in science.*** **Paper presented at OERI conference "The Promise and Perils of Alternative Assessment," Washington, DC.**

Shepard, L. A. (1988, April). ***The harm of measurement-drive instruction.*** **Paper presented at the annual meeting of the American Educational Research Association, New Orleans.**

Shepard, L. A. (1991). Psychometricians' beliefs about learning. ***Educational Researcher, 20*****(6), 2-16.**

Shepard, L.A., & Dougherty, K. C. (1991, April). ***Effects of high stakes testing on instruction.*** **Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.**

Shepard, L. A., & Kreitzer, A. E. (1987a, April). *The Texas teacher test.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Shepard, L. A., & Kreitzer, A. E. (1987b). The Texas teacher test. *The Education Researcher, 16*(6), 22-31.

Solmon, L. C., & Alkin, M. C. (Eds.). (1983). *The costs of evaluation.* Beverly Hills, CA: Sage Publications.

Solmon, L. C., & Fagnano, C. L (1990). Speculations on the benefits of a large scale teacher assessment programs: How 78 million dollars can be considered a mere pittance. *Journal of Education Finance, 16*(1), 21-36.