ESTIMATING THE COSTS AND BENEFITS OF LARGE-SCALE ASSESSMENTS: LESSONS FROM RECENT RESEARCH

CSE Technical Report 319

James S. Catterall

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

November 1990

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This report is forthcoming as an article in *Journal of Education Finance.*

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

The post Nation at Risk years have brought pervasive reforms costing billions of dollars to public education, an enterprise already receiving more than half of state and local budget appropriations. Financial responsibility for elementary and secondary schools has shifted steadily over two decades to the state capitals, where legislators must defend school appropriations to taxpaying voters as well as to constituent groups pressing for other services. In this environment, is it any wonder that large-scale accountability systems such as state educational assessments are a growth industry?

Each of the following questions can be linked to statewide or national testing efforts: Are contemporary youngsters out-performing those who endured the indicted school systems of the 1970s and early 1980s? Are American high schoolers catching up to the supercharged youth of Japan in mathematics and science? Are children learning the core of things that all should know before they leave high school? Will California schools ever overtake those in New York? Are our teachers fit to teach?

When we ask such questions at the state or national level, as Americans are quite bent on doing nowadays, the answers are sought through large-scale assessment strategies; this 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. Since these student testing programs and teacher certification tests are large in scale they can cost sponsors millions of dollars, as well as induce a variety of less visible costs. A basic question underlying this discussion is just what do we receive in return for these hefty assessment outlays?

Overview

This paper does not report a "cost-benefit analysis" of a large-scale assessment activity *per se*, nor does it attempt to evaluate the worth of the species in any global way. Its purpose is to present and substantiate some critical lessons and observations concerning (a) how the costs and benefits of these assessments can be appraised, and (b) the utility of these lines of research to policymakers and educators. While these questions are not new to the education research community (see Solmon & Alkin, 1983), the paper does bring some new data to bear on them. One source of insights is a recent study of the Texas teacher certification test (Shepard & Kreitzer, 1987a; 1987b). Another is our national study of minimum competency testing (Catterall, 1989).

Eight lessons are identified and briefly discussed. Taken together, these lessons point to several overriding implications:

1. Although the large-scale assessment-as-intervention is not well tailored for the classic cost-benefit and cost-effectiveness analysis models advanced by economists, key insights to these activities can be generated by cost-benefit analysis approaches.

2. The costs and benefits of large-scale assessments are generalizable neither to large-scale assessment activities as a class nor to particular types of these activities, such as the teacher certification test. Instead, the resource demands and products associated with these testing systems are dependent on the specific ingredients needed for their administration and on the uses made of the resulting information.

3. There appear to be credible political explanations for the evident lack of policymaker interest in such cost-benefit studies. The arguments here imply that resource allocations to large-scale assessments are likely to be neither educationally nor economically optimal. Instead, we should expect such allocations to be politically optimal.

Lessons From Research on Large-Scale Assessments

We turn now to the principal arguments of the paper, presented as eight observations or "lessons" drawn from cost-benefit research on large-scale assessments. Before proceeding, readers should note that the observations do not result from an accumulated weight of in-depth cost-benefit type studies, since no such weight has been registered. The points tend to build on the small number of interesting developments reported (particularly Shepard & Kreitzer, 1987a, 1987b; Solmon & Fagnano, in press), as well as on the author's experiences in conducting cost-benefit type analyses of educational assessment practices (Catterall, 1984, 1989). We also base inferences on the paucity of research itself. The lessons have practical importance for would be cost-benefit researchers, either as constructive working suggestions or as caveats warranting consideration.

Lesson 1: The classic cost-benefit analysis model cannot underwrite a comprehensive analysis of large-scale assessments.[1]

Most education researchers and education leaders have little exposure to formal training in economics or cost-benefit analysis and are glad for this. One consequence is the generally creative ways that the terms cost-benefit analysis, benefit-cost analysis, and cost-effectiveness enter educational policy discussions. Economists apply strict definitions to these ideas, and these definitions will be used here to provide common ground for the discussion.

Cost-benefit analysis (CBA), a term interchangeable with benefit-cost analysis, refers to the comparison of the costs of an enterprise to the benefits of that enterprise where costs and benefits are measured in monetary terms. CBA requires the analyst to estimate dollar equivalents for costs and benefits of a program under scrutiny. This facilitates clean analytical comparisons, such as ratios of benefits to costs, the amounts that benefits exceed or fall short of costs, and the implied rate of return on dollars invested in activities (Levin, 1983).

We note extreme convenience in achieving dollar quantities for costs and benefits of public programs. In this desirable state of analytical affairs, 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. 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.

Is CBA useful for large-scale assessments? When we try to apply CBA principles to large-scale assessments, 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 extracted from the reports of Shepard and Kreitzer (1987a, 1987b) points to these difficulties and is shown in Table I:

---

[1] Rather than providing a glossary, this discussion will define its key terms as it proceeds. We begin Lesson 1 by establishing the meaning of cost-benefit analysis.

Table I

Example Effects of Teacher Certification Tests

Decertifying unfit teachers;
Empowering unfit teachers who manage to pass;
Decertifying disproportionate shares of shop teachers, special education teachers, and non-academic personnel;
Unjustly failing some minority teachers;
Altering teacher morale;
Altering public's appraisal of the teaching force;
Altering education system's relations with the legislature;
Raising educational achievement levels.

The nature of man of the possible effects of teacher certification tests shown above (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 the costs of certification tests, we are unlikely to find an ideal answer. That is, we cannot hope to say that the dollar values yielded by a competency or teacher test exceed or fall short of the costs of putting on the show. 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.

Lesson 2: Cost benefit analysis principles can be applied to a subset of all effects and thus contribute to an appraisal of the worth of large-scale assessments.

The previous section argued against reliance on cost-benefit analysis techniques to appraise the overall worth of large-scale assessments. This disclaimer results from the non-monetary nature of some portion of the benefits. 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 monetarized and examined *vis a vis* assessment costs. Under some circumstances, the results can be revealing and decision-relevant.

One example of this is Shepard and Kreitzer's (1987a) analysis of the recently administered Texas teacher test, the TECAT.[2] This test, given to more than

[2] We secured two versions of this paper, noted in the references. The detailed discussion of costs and benefits appears in Shepard and Kreitzer's (1987a) paper. The *Education Researcher* version has a truncated discussion of costs and benefits. This may attest to the substance of our comments concerning the receptivity of the education research community to economic models. It may simply be a matter of *ER's* space limitations.

200,000 practicing teachers, was ostensibly designed with a simple set of objectives in mind: to ensure that teachers possessed 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. They applied cost-benefit analysis principles to some of these.

One such instance was comparing the costs of the assessment to the recaptured salaries of ousted teachers. The total public cost of implementing the TECAT was about $36 million, amounting to nearly $30,000 per failed teacher. (Additional costs for failed teacher unemployment, welfare, and retraining programs were not considered.) One "benefit" of removing these teachers was that their salaries totalling $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 probably would have shown present value savings in the hundreds of millions of dollars.[3] This perspective would have suggested a very high ratio of benefits to costs on this one dimension alone.

Another CBA example from this same study was its inquiry into the opportunity costs of spending some thirty million dollars 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 other alternatives. 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.

Another example of a partial cost-benefit analysis is found in Solmon & Fagnano (in press). This analysis suggests that unfit teachers lead to undereducation of pupils and to lowered educational attainment. This behavior leads to lower productivity and to social costs which can be estimated in dollar values. An undeveloped but analogous line of reasoning applies to our recent research on minimum competency tests; we found strong indications that tests required for high school diplomas 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 the authors do not claim to provide monetary estimates of the total benefits deriving from the tests under scrutiny. Rather, they adopt a selected focus and proceed to produce costs and benefits in dollar terms that apply to their focus. This evidence may then play a contributing role in policy discussions concerning the merits of such assessments, an implicit goal of this sort of research. A partial analysis may provide a very compelling argument for supporting or abolishing a large-scale assessment where a benefit or cost is shown to be preemptive.

Lesson 3: The classic cost-effectiveness analysis (CEA) model is suited to a comprehensive analysis of large-scale assessments, but the nature of large-scale assessments limits the use of CEA in practice.

---

[3] It is not our intention in this discussion to overly dwell on technical cost analysis issues. Discounted present value techniques can be used to translate streams of costs, such as salary outlays, to a single equivalent lump sum.

As illustrated in the first two lessons, limitations in assigning monetary values to benefits can 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 exemplary finding from a cost-effectiveness approach was mentioned above: the TECAT costing $36 million decertified 1199 teachers; the cost per unit of effect would read $29,703 per teacher expelled.

Cost-effectiveness analysis requires that effects be quantified, but not that these quantities be in dollars. Relaxing the need to provide monetary estimates can bring monetarily problematic benefits 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,[4] 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.[5]

The basic mechanism of linking costs and effects in cost-effectiveness analyses is the calculation of ratios between units of outcome and the dollars expended to achieve them. Examples include the $29,703 per failed teacher above; others could be a ten point increase in a scaled measure of public confidence in the schools resulting from the institution of a test costing $10 million, each point gained thus requiring $1 million; or a study showing two percentage point gains in SAT math scores for each $5 million expended for diagnostic testing in grade 9.

While comparing effects in their natural units to costs in dollars avoids the limitations described for CBA, the tendency of large-scale assessments to have multiple effects limits the utility of CEA to decisionmakers. 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 percent decrease in teacher morale, Y units increase in reading achievement, and Z percent of special education teachers decertified. Assuming that these effects can be measured satisfactorily, this basket will portray what the assessors gained for their money. But this information does not particularly 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. Effects estimates in their natural units do not provide a ready answer to the question of whether the assessment was worth it. We may say that we gained two apples and three oranges and lost one plum, but whether this is worth the dollar spent must yield to human judgment.

A second shortcoming of CEA approaches is the difficulty of comparing baskets of anticipated effects where alternative assessment strategies are

---

[4] The analysis by Shepard and Kreitzer notes various uncertainties involved in assessing the number of teachers forced out by the TECAT. For example, some teachers decided not to be tested and retired or left the state. But normal retirements and attrition were also expected. Estimating the number of pushouts from this sort of test is thus not a simple matter of counting those who fail to pass the test.

[5] Lesson 4 discusses some of the evaluation issues implied by this listing.

contemplated. The importance of such a decision context cannot be overstated, since choosing alternative means to a desired set of ends lies at the heart of public policymaking, including deciding on large-scale assessment policies. It is likely that the baskets of effects associated with alternative 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 the assessment strategies clear. This case assumes agreement on which effects are liked and which are disliked. CEA methods simply do not provide any assistance with the problem of comparing, for example, the decertification of 100 incompetent teachers on the one hand with a 0.13 standard deviation reduction in teaching force morale on the other. We could make such a comparison in the idealized world of CBA, where dollars are assigned to all effects.[6]

Thus it appears that while cost-effectiveness analysis approaches allow us to sidestep our inability to construct comprehensive dollars-only cost-benefit analyses, available CEA techniques have their own difficulties where the provision of policymaking advice is concerned. Yet CEA and CBA perspectives have been shown to offer potentially useful information to the policymaker. 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.

Lesson 4: The traditional problems of inferring program effects in complex systems affect CBA and CEA studies of large-scale assessments.

Attributing general educational or social effects to small contributors within large complex systems is a challenging analytical prospect. Such is the task implied when we list among the outcomes of large-scale assessments effects such as pupil learning, teacher morale, public confidence in education, and education system relations with the legislature. When we hope to quantify such system-wide effects either in naturally occurring units (CEA) or in dollar equivalents (CBA), the first order of business is the inherent research or evaluation problem. Distal outcomes as these must be accounted for in the context of a full model of influencing factors. Research that can convincingly partial-out the effects of a mere educational assessment on any of these outcomes would have to be clever and would probably be expensive.

One approach is that taken by Shepard and Kreitzer (1987a, 1987b) in the reports noted above. They employed a roughly 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 approach appears to have made productive use of the resources available for the study. It had the advantages of tapping into stakeholders for whom effects were important and tangible. It had disadvantages in the form of perception-based measures, possible inaccurate recall, and possible response bias. On the last point, respondents may have surreptitiously favored or disfavored the TECAT for unaccounted reasons, and thus reformulated their appraisals of past events.

---

[6] Levin (1983) suggests a cost utility analysis procedure that provides a judgmental model for assessing baskets of incommensurable effects. This model requires empirical assessment and scaling of how much different outcomes are valued by parties relevant to a decision. Also, as discussed in a subsequent section of the paper, political processes implicitly weigh and decide these trade-offs through voter behavior and the delegation of decisionmaking authority to officials who then juggle competing interests. Seldom, however, are the considerations of trade-offs through these processes very explicit. In the case of large-scale assessments, the frequency of explicit treatments reported so far is zero.

A general approach to assessing the effects of an assessment system would be quasi-experimental research. Two sorts of designs are potentially applicable: (a) Longitudinal tracking of changes within a system and (b) cross-system comparisons.

In the first type of design, changes accompanying the implementation of assessment might be measured longitudinally. Steps would be taken to account for the effects of other influencing factors on the system-wide outcomes of an assessment system. The analyst may hope or assume that other determinants of the outcomes will remain constant or neutral in effect over the relevant time period. Unfortunately, these crucial conditions are difficult to assess, and in the case of broad outcomes such as pupil achievement, an assumption that other influencing conditions do not change from year to year may be untenable.

In the second type of design, differences across educational systems might reveal the influence of a large-scale assessment mounted in one of them. One research strategy is comparing a state with a teacher certification test to one without; another is pooling states for multivariate analysis. These designs pose challenges to the researcher. It is unlikely that differences in assessment systems (e.g., presence in one state versus absence in another) would be the only factor accounting for outcomes of interest; if early teacher retirements were under scrutiny, influences such as salary and benefit structures, working conditions, and alternative work opportunities must be considered. Attributing effects to a stiff teacher assessment system alone cannot be justified.

In a multivariate design, it may be impossible to identify sufficient cases to sustain an analysis; "large-scale" has generally referred to state-level assessments in this discussion, and this may limit the population to 51. Such a sample size would allow for a very constricted set of determinants of state-level outcomes in question, such as pupil achievement or motivation; thus the resulting models would probably be very poorly specified. Of course, there are alternative frames of analysis for which studies would not be limited to such restricted N's. For example, assessment practices, costs, and effects might be examined using a state's 250 school districts as units of analysis, thereby affording more power to resolve test effects and their relationships to costs.

Perspectives on the substantial research needed to appraise state-level effects of large-scale assessments are not particularly optimistic. Analysts have relied largely on perception-based single case studies to attain effects measures. This prospect may incline the reader to retreat to the more selective, partial cost-benefit or cost-effectiveness 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 decisionmakers. 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.

Lesson 5: Commonly held assumptions regarding educational assessment promote widespread beliefs that certain benefits will follow from its practice.

Assessment is a conspicuous and regular fact of educational life, and tests have become an accepted part of American culture. The grumbling of the measurement profession notwithstanding, educators, policymakers, and lay citizens alike generally believe that tests are capable of measuring the knowledge and skills held by individuals. Thus the use of tests to assign individuals, such as students and teachers, to particular knowledge-based or skill-based statuses is legitimate; such awards as 10th grade standing, a diploma, a teaching post, or a pink slip come to mind. When we add high stakes to the accepted virtues of tests, we can assume that the prospect of being tested will induce desired behaviors in those to be tested.

Consequences widely expected from a test that can change one's status for better or worse include studying harder, attending and being more attentive in school, and correcting anticipated or revealed deficits.

It follows that policymakers and citizens can easily assume that tests will infallibly reap certain benefits. The tests used for large-scale assessments of students and teachers are no exception. Teacher certification tests will cause teachers to brush up on their skills, and because they can be designed to assess critical basic skills, they will point to those who are deficient in this domain. Tests required of pupils for high school graduation will cause students to tend to business, and 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 of the assessments themselves that might pose challenges to them, a corollary to Lesson 5.

Data-free assumptions of benefits are actually made by testing policymakers and educational leaders, which 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 undertaken in more than half the states. The work examined in great depth the testing practices in four selected states and included in-class surveys and discussions with 736 ninth and eleventh graders.

The most graphic testimony revealing tendencies to suppress inquiry regarding assessments is the dismal state of policy-relevant information concerning pupil performance on the tests we probed. In our national study, we found not a single school, district, or state which 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 the 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 decisionmaking 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 current research 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 high schools we studied knew that passing a competency test was required for their graduation. About 45 percent of 9th graders knew of this requirement, a figure that grew only to 58 percent for 11th graders.

These perhaps puzzling levels of student awareness stand in stark contrast to the assumptions held by our respondents, whose schools universally required such tests and administered them for the first time in the 9th or 10th grade. More than 400 of the 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 have

tentatively 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.

Another touted benefit of pupil competency tests is their role in the aeromedicine of student skills. In our study, educators universally heralded the capacity of competency tests to spot individual learning difficulties and point 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 percent of students in general and 54 percent of test failers themselves answered yes. Students with low grades (those closer to the behavior in question) were less likely to offer an affirmative response to this remediation question than students with high grades.

In a larger sense, our research provides indications that educator and policymaker 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.

Lesson 6: The benefits of large-scale assessments will depend heavily on the uses made of assessment information.

The introduction noted that the benefits of large-scale assessments may not accrue to such devices generally nor to any selected type, such as the minimum competency test. 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 occurs because educators and the education system use assessment information in different ways.

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 gulf 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." In sharp contrast, other schools reported establishing special remedial classes for test failers and other concerned attempts 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 uneven.

This reasoning toward benefit-variability can be taken a step further. Remedial classes spawned by competency testing on the one hand may 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.

Lesson 7: The costs of large-scale assessments typically far exceed their published budgets.

This is a simple but crucial lesson. 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, the reported appropriation was $4.8 million. By the time researchers Shepard and Kreitzer (1987b) turned off their adding machines, the public costs of the program were more like $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 support the administration of 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 inservice training 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. Induced private costs, such as teacher study time and privately paid workshops, are also appropriate candidates for inclusion in a complete cost accounting.

A similar but smaller-scale example appears in our study of a school district's pupil information system. In this case, the budgeted costs for curriculum-matched testing system in the elementary schools amounted to 80 cents per pupil, but the costs of all ingredients identified as necessary to its operation totalled $34 per pupil (Catterall, 1984).

The primary reason that this "ingredients" perspective is appropriate to a cost appraisal of large-scale assessments is that the productive value of these ingredients for alternative uses represents the opportunity cost 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. Such an alternative endeavor is sacrificed where an assessment program is maintained, and this sacrifice represents its true costs.

An additional line of argument is proposed supporting the observation that the true costs of large-scale assessments inevitably exceed their published budgets. This is the possibility, if not the likelihood, that negative effects will accompany large-scale assessments and that these effects will induce real costs. Analogous to the Solmon and Fagnano (in press) suggestion 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.[7]

If competency tests tend to push out youngsters who might have persisted and benefitted otherwise, some of the costs of dropping out could be pinned on such tests. This accusation seems particularly justified in some of the cases cited above, where the test seems to be used in a degrading fashion 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

[7] In a fully specified model for self-reported chances of finishing high school by 736 ninth and eleventh graders in four states, failing a required test for graduation was a strong and significant independent predictor. The model controlled for student background, performance in school, and various school context conditions (Catterall, 1989).

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 lost tax collections and a 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.[8] This in turn could be "costed" according to the procedures used in our cost-of-dropouts analysis.

Lesson 8: Cost-benefit analyses of large-scale assessments have the potential to advance educational or economic optima. Large-scale assessment policy decisionmaking may be likely to seek a political optimum instead.

This final lesson points toward an important conclusion of the discussion. The objective of cost-benefit or cost-effectiveness analyses of large-scale assessments is the attainment of economic or educational optima. 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 CBA and CEA, however, the analyses have similar implicit purposes: showing policymakers just what they are spending 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. These are classic rational policymaking perspectives.

The Triumph of Politics?

On the basis of the lessons argued above, it appears that a perspective alternative to rational policymaking has governed large-scale assessment decisions: a political perspective. Goals other than educational or economic values are pursued by decisionmakers deciding to sponsor or administer large-scale assessments. For state legislators, frequent arbiters of large-scale assessments, these goals may center on creating favorable impressions across the electorate and among constituencies. Legislators will naturally desire to show that they are in command of crucial educational and social issues, or at least that they are responsive to public outcry. Similar motivations can be ascribed to education leaders.

Before an elaboration is offered, readers should note that the discussion reaches beyond its immediate data to suggest this line of argument. No study of legislators or assessment decisionmakers regarding their political motivations was undertaken for this work. Nonetheless, available literature on the pupil competency testing movement (e.g., Jaeger & Tittle, 1980) notes the political roots of such assessments. We find additional support for a political thesis in the arguments presented above. The central point is the general absence of cost-benefit information in the assessment policy arena despite its apparent feasibility of

---

[8] Various studies of the determinants of school dropout using High School and Beyond data incorporate student-expressed chances of finishing school as a predictor variable. For example: Stern, D., Catterall, J.S., Alhadeff, C., & Ash, M. (1986). *Reducing the high school dropout rate in California.* Berkeley, CA: Institute for Governmental Studies. Also: Eckstrom, R.B., Goertz, M.E., Pollack, J.M., & Rock, D.A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers College Record, 87*(3), 356-373.

attainment. This suggests a paucity of real demand for such information by decisionmakers. This reasoning and its implications are now explored briefly.

Absence of studies. The shortage of relevant studies is noteworthy. Only trace quantities of cost-benefit research regarding the tests we have discussed have been reported. This dearth of analyses which synthesize 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 decisionmaking, is practically nonexistent. Our knowledge of the effects or benefits of large-scale assessments can only be described as thin. Decisionmakers and educators seem rather content with limited-data impressions or their preconceptions about large-scale assessment effects. These are certainly inexpensive and undemanding sorts of information, at least in the short run. This circumstance also implies that other considerations probably drive assessment policy decisions, a point revisited in the discussion below.

Feasibility of cost-benefit type studies. We argued that 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 novel 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.

Little evident demand for cost-benefit studies. There is little apparent demand for economic or educational information regarding large-scale assessments on the part of policymakers or education leaders. We can safely assume that if such information were desired, such studies would be commissioned and reported. To the contrary, we have suggested that such information may be distinctly unwelcome. For legislators, a cost-benefit study might reveal that a particular large-scale assessment was the educational equivalent of the Sargent York Gun, the $5 billion U.S. Army project that never worked. From the legislator's point of view, why spend the resources for such a study if it might serve to undermine a program deemed politically beneficial? This plausible case suggests a situation where the benefits and costs privately considered by legislators might differ from the benefits and costs usable in their public expressions regarding sponsored programs. A resulting hesitancy to commission cost-benefit or cost-effectiveness studies is understandable.

Discussion

A distinct suggestion of our analysis is that policy decisions concerning large-scale assessments have been made on some basis other than rational (or informed) cost-benefit or cost-effectiveness thinking. Politics looms as a strong contender.

The foundations of large-scale assessment decisionmaking might be seen in the environments that spawned large-scale assessment movements in the first place. One of our examples, the minimum competency test, emerged in the mid 1970s in response to public and employer concerns that the high school diploma lacked meaning (Jaeger & Tittle, 1980). As the skill levels of newly graduated high schoolers became a major public issue, many state legislatures crafted responses in the form of accountability assessments. If a test could be designed and administered, proceed to flunk a few students and not otherwise cause too many difficulties in the field, the system would be scored a political victory by sponsors. As a result, legislators wishing to claim they had tackled an important problem would add

feathers to their caps. The roots of teacher certification tests seem directly analogous.

Unless someone demonstrates a superlative, costly, or egregiously inappropriate consequence, what else occurs as a part of the assessment bargain, particularly effects of an educational or economic sort, seems to have no great interest or importance. Whatever may have been demonstrated as the effects of graduation or teacher testing so far, no operational large-scale assessment systems in the United States have ever been abolished, at least to our knowledge. The sponsors of more than 25 state graduation testing systems (and of scores of additional state educational assessments not discussed in this paper) thus argue that the benefits of these tests warrant their costs.

An interesting implication of this political perspective on large-scale assessments is that most of these tests would probably survive even if their recognized benefits were far less than their acknowledged costs. Just as the benefit to a sponsoring legislator may accrue through the act of legislating a large-scale assessment, the act of repealing a large-scale assessment may be regarded as simply too costly to endure. A decision to repeal might be interpreted publicly as a retreat in the battle to instill standards for pupil and teacher performance. Thus the assessment system would be supported even if it proved at heart to be a losing human resource proposition.

Summary

To recap briefly, the discussion first illustrated how the costs and benefits of large-scale assessments can be approached by the policy analyst. We noted that both cost-benefit analysis and cost effectiveness analysis have applications to large-scale assessments. Their respective limitations were also catalogued. CBA suffers by requiring monetary values for all benefits. This leaves many outcomes of assessment ineligible for analysis, or at best subject to questionable monetarizing tactics. In contrast, CEA suffers from problems of weighing incommensurate outcomes; the results of CEA may not be able to guide decisions where ambiguous mixes of effects are anticipated. Because of these factors, and because some of the hoped-for global outcomes of large-scale assessments are difficult to detect with precision, narrower partial cost-benefit studies were recommended. Illustrative examples were offered, and their general lack in the literature was noted.

We also argued that the benefits and costs of large-scale assessments cannot be generalized to specific types of these assessments in principle. Instead, these values are dependent on the ingredients required for the maintenance of particular large-scale assessments and on the actual uses of assessment information. We illustrated these points using our previous research which revealed that there appear to be beneficial and malevolent strains of effects in pupil competency testing systems. This argument suggests that cost-benefit findings for a given assessment system must be examined critically before they are attributed to other systems.

We reported that the costs of large-scale assessments typically (and importantly) far exceed the direct allocations made to support such systems. Decisionmaking regarding large-scale assessments should attend to the true opportunity costs of these systems and not merely to the initial appropriations. The latter sums may approximate seed money when final cost totals are estimated.

Finally, we argued that resource allocations to large-scale assessments appear to be decided in the relative absence of explicit cost-benefit analysis findings, for a variety of reasons. Technical impediments, though some exist, do not seem to fully account for the lack of such research. We speculated on the apparent lack of

interest in, and even distaste for, cost-benefit findings among legislators and educators. Our observations suggest that large-scale assessment policy decisions seek maximization of political values. They do not attest to overriding concerns for economic or educational maxima.

References

Catterall, J.S. (1984). *The costs of instructional information systems: Results from two study districts* (Project report). Los Angeles: UCLA Center for the Study of Evaluation. Also paper presented at the 1984 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.

Eckstrom, R.B., Goertz, M.E., Pollack, J.M., & Rock, D.A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers College Record, 87*(3), 356-373.

Jaeger, R.M., & Tittle, C. (Eds.). (1980). *Minimum competency achievement testing.* Berkeley, CA: McCutcheon.

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

Shepard, L.A., & Kreitzer, A.E. (1987a, April). *The Texas teacher test.* Paper presented at the 1988 Annual Meeting of the American Educational Research Association, Washington, D.C.

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.

Solmon, L.C., & Fagnano, C.L. (in press). Speculations on the benefits of large scale teacher assessment programs, or how 78 million dollars can be considered a mere pittance. *Journal of Education Finance.*

Stern, D., Catterall, J.S., Alhadeff, C., & Ash, M. (1986). *Reducing the high school dropout rate in California.* Berkeley, CA: Institute for Governmental Studies.