TESTING AND EXCELLENCE IN EDUCATION: SOME QUESTIONS ABOUT COSTS AND BENEFITS

Eva L. Baker
Joan L. Herman
James S. Catterall

CSE Report No. 223
1984

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

Introduction

Attention to issues of excellence in education has been heaped upon the public in recent months. Critics with varying prestige, information, and insight have reminded us with resounding harmony of this neglected purpose of American schools. What is unusual this time, however, is not this concern (for we have worried about educational quality all along), nor even the connection of schooling to societal productivity and survival (for we are unshakable in this belief). What is different now is the salience of cost considerations in both the larger debate and also in propositions offered for action. Will our society back-up a desire for markedly better schools with resources? Who will pay for the sorts of things our recent commissions suggest are needed, such as providing monetary incentives for quality teaching, enhancing the general attractiveness of the educating professions, adding solid requirements to high school curricula, or bidding science and math talent away from industry?

Testing in the schools holds a curious center spot on this stage -- as the principal harbinger of the problems of education, as a means to their solution, and as an activity commanding important resources. First, the decline of educational quality has no more common expression than in the performance of youngsters on nationally recognized tests. Declines in SAT scores and sagging performances on statewide assessments embody current indictments of the schools. Second, tests have been introduced as levers for improvement. Teachers who pass tests may be more effective in the classroom; high schoolers facing test hurdles for graduation may at last shape-up; more and better

testing in the lower grades may keep potential failures and their mentors on roads to success. And finally, testing is itself costly, and its proliferation has caused us to think more systematically about what we are getting for what sort of investment in our rush to assess.

In this discussion, we probe questions about our commitments to testing. We briefly describe the explosion of tests as a forerunner to the current embroglio over educational excellence. We then raise important questions about what the results of tests mean and suggest that they may reliably convey less than their sponsors or users might hope. And we report on our recent inquiries into the costs of testing -- inquiries spawned by our suspicions that sizeable resources fuel current assessment efforts in the schools. Just where does testing stand in a world of competitive demands for public and educational resources?

Testing as Standard Bearer

Slightly ahead of our pandemic interest in excellence was the recognition by legislators and school boards that some quality measure should be routinely available to judge the progress of our students' achievement. Competency testing for children in schools has been in place for approximately five years on a large scale basis. The option for such testing first placed educators in the unhappy situation of trading off local control (and locally appropriate measurement) for the demands of comparability among students, schools, and school districts. In addition, educators were placed squarely in the midst of boring, recurrent, but important contention about whether there

were tests good enough, precise enough, valid enough, fair enough, and cheap enough to regularly employ and to trust as measures of school output. Certain localities had earlier and stronger confidence in the quality of tests and measures to assess student performance. In those cases, individual states developed, administered, and defended in court statewide measures of achievement. In some cases, passing these tests was critical, for ultimate test failure also meant failing to obtain a high school diploma, a reward previously available to anyone who sat in classrooms on a moderately regular basis. Other states and city school districts used achievement testing in either more relaxed or more stringent ways. Certain states now collect information on children's performance by sampling students or results on specific test items, or both. Sampling information gives a cross-section of performance at a given time for all students, but provides no specific information about a given student's achievement. Some schools use test results for actual decisions related to the progress of individual students, such as when test scores are considered as an important criterion when children move from one grade level to the next.

Regardless of format, these tests are intimately lodged in the present debate about educational excellence and productivity. We will now consider issues of test meaning, test costs, and test uses so that the practical utility of testing as a mechanism for improving educational quality can be judged. We will exhibit throughout some skepticism about the quick fix tests provide to the problems of education.

Test Meaning

What do test scores mean? As sophistication in achievement testing has developed, so have doubts concerning what we can say about the importance of any single test score. Certainly, there is less uncritical acceptance of test scores as serving as infallable measures of intelligence, aptitude, or ability than ever before. Yet, a "good" score is sought by all, whether on the Scholastic Aptitude Test or an a Reader's Digest word power quiz. Test score meaning functions psychologically in the same ways as good grades in school. We know some teachers are arbitrary or casual about how they assign grades and that the absolute meaning of marks of A or B- is suspect. But we want to do as well, or better, than our cohorts, even on unreliable measures. Test meaning is normative (who is better than whom). Unfortunately that comparative base is used in achievement test-based inferences of educational quality. Most simply, if schools (or students) are arrayed from best to worst on the basis of test scores, it is possible that the distance between very good students and schools and poor students and schools in terms of actual performance might be small. Nonetheless, half of the schools (and students) will be definitionally below average.

Even if we could magically pop ourselves out of a best-to-worst framework, test meaning still has seemingly insuperable problems. Tests are valid (that is accurate and appropriate) indicators of student and school performance only under special circumstances. First, the test needs to reflect the actual program of study children receive. If an examination tests content and skills at odds with

school experiences, then the information it provides is very weak. Based on a number of analyses, for example, those by Walker and Schaffarzick (1974), Floden, et al (1980), Herman and Cabello (1982), much of what presumably is taught (as reflected in the principal textbooks children use) and that which is measured on either state developed or commercially available tests do not match up. Targetting both instruction and testing on the same agreed upon goals and objectives seems a simple solution, but one that is made a difficult proposition when various commercial publishing houses control much of the content of the curriculum.

Second, test meaning is also affected by when the measurement is made. Testing that takes place once and only once a year provides a cross-section of performance at an administratively convenient time and invariably overlooks gains made by students whose learning patterns are a little delayed.

Third, test meaning is dramatically influenced by seemingly irrelevant attributes of the test. One of the most painful examples is the use of multiple choice tests asking students to chose among better or worse worded paragraphs, an indicator serving as a proxy measure for written composition. Any of us who has struggled with the task of thinking clearly enough to write something does not need a cognitive psychologist to tell us that very different thinking patterns are required to produce something rather than to correct it. In fact, that is why we have editors as well as original writers. Thus, the format of tests, usually chosen for ease of scoring, can mislead our understanding of student achievement.

In addition, we have the rather peculiar effect of particular content. Even when our stated goal is to assess reading, or writing, or science concepts, the particular idea chosen to include on the test, e.g., whether the poem is by Blake or Donne, critically affects our ability to perform. Evidence in the field of written composition mounts in support of this idea; the effect of topic is dramatic (particularly when topic knowledge is not the main purpose of the test).

Finally, and most significantly, the problem of test quality has not been solved. Test quality involves simply doing a good job of test preparation and using sensible scoring standards. But simply reading many of the extant commercial or locally prepared tests routinely given to students will convince anyone that clear, right answers are not always available, even for experts, and that it is possible to read and understand test questions in a variety of legitimate ways, only one of which matches what was in the test writer's head.

Test Costs

Now that we have cursorily explored the weaknesses of existing tests, let's take a look at what they cost. The UCLA Center for the Study of Evaluation (CSE), with support from the National Institute of Education, has been conducting a national study of testing (Baker, 1983; Dorr-Bremme et al, 1983; Herman & Dorr-Bremme, 1983; Burry et al, 1982). The primary purpose of this study is to find out how much testing is occurring, what kinds of tests are in use, what use teachers make of test results, and the costs of testing. Our approach

included intensive interviews as well as a national survey of 100 school districts representing geographic regions of the USA, various degrees of urbanization, poverty, and wealth, in addition to grade level and subjects taught by teachers.

In our effort to capture the costs of testing, we conducted a close-up study in two school districts. Within these districts we attempted to estimate the magnitude of costs of various testing activities either in terms of their primary units (such as teacher or counselor hours devoted to testing) or direct dollar costs of materials and services purchased, and then to convert all resource estimates to dollar equivalents.

District and school personnel in each district were interviewed to identify the types of tests administered; the nature of test administration, analysis and dissemination activities at the district, school, and classroom levels; the types of materials and services purchased from outside vendors; and the amount of personnel time devoted in each phase of test implementation. Let's consider the results for one of the districts, a large urban district labeled Metro for the purposes of reporting. It exemplifies the district which administers tests to comply with external demands for accountability.

Accountability Costs

Metro has a heterogeneous student population numbering over 500,000. Its over 600 schools are staffed by approximately 17,000 classroom teachers, supported by a budget that includes a complex mix of local, state, and federal funding. Cost per pupil spending is about $1,900, and aggregating over all available sources approximately $3,300 could be spent on the most impacted student.

What is Metro's testing program? Metro administers a variety of basic skills tests for various internal and external purposes. Most dominant among these tests is the Continuum-Based Skills Survey, which is administered yearly to all students in grades 1 through 6 to provide the district and teachers with information on students' progress through and mastery of the mandated district curriculum. The test also fulfills reporting requirements for state and federal programs, although the district continues to administer norm-referenced tests at grades 3 and 5 to satisfy these same requirements and for its own evaluation purposes.

Proficiency tests in reading, math, and writing at grades 7 and 10 (plus retakes as necessary) is another major district activity. These tests were initiated in response to a state mandate requiring minimum competency assessment for high school graduation. As at the elementary school level, standardized norm-referenced tests are administered to satisfy federal program requirements for students participating in such programs, and are administered to all students at grade 8 for purposes of program evaluation. A mandatory state assessment test administered at grades 1, 3, 6, and 12 is a final component in the district testing program.

Metro's basic skills testing program is summarized in Table 1.

Table 1

SUMMARY OF METRO DISTRICT BASIC SKILLS TESTING

School Level	Tests	Grades	Type	Purpose
Elementary	Continuum Tests	1-6	Criterion-referenced	Pupil diagnosis, curriculum planning, 3-5; Chapter I reports to State/Fed
	Norm-Referenced Test	3,5 (6 optional)	Norm-referenced	Instructional program assessment
	Norm-Referenced Test (Spanish)	1-6	Norm-referenced	Individual tests for all children receiving Spanish reading instruction
	State Assessment	entry, 1,3,6		State Assessment
Junior High	Math Proficiency	7 plus retakes	Proficiency	Pupil progress, math
	Writing Proficiency	7 plus retakes	Proficiency	Pupil progress, language, writing
	Reading	7 plus retakes	Proficiency	Pupil progress, reading
	Norm-Referenced	8	Norm-referenced	Instructional program assessment
	Norm-referenced (Chapter I schools)	7,8,9	Norm-referenced	State/Federal reports
Senior High	Math Proficiency	10 plus retakes	Proficiency	H.S. graduation requirement, math
	Writing Proficiency	10 plus retakes	Proficiency	H.S. graduation requirement, writing, language
	Reading	10 plus retakes	Proficiency	H.S. graduation requirement, reading
	Standardized tests (Chapter I schools)	10-12	Norm-referenced	State/Federal reports (10 out of 49 high schools)

What are the costs for Metro's testing program? Metro assigns significant central office resources to its basic skills testing program. The district Research and Evaluation Unit houses five professionals and five clerical staff who work exclusively with the district tests -- scheduling tests and test-related activity, arranging purchase, delivery, and scoring of test materials, writing reports of results, etc. The unit also maintains two automated scoring machines which require four to six operators when tests are being scored, and requires about two full time equivalent computer programmer/consultants to assist in its information processing.

In addition to these central office costs, Metro also incurs cost through a variety of services and purchases outside the central offices. Among these costs are the operation of scoring centers in the districts' 10 regional offices, contracts for test processing, supply costs for all tests, and a long term contract with an outside laboratory for the development of its elementary skills assessment test. (Total spent on this contract alone since 1976 is about $1,000,000.)

Summing across all these expenditures, Metro spends $1,436,000 at the district level for its testing program, representing an average cost of $2.64 per pupil (see Table 2), if tests were equally distributed across students, which they are not. The costs of specific tests within the district program, prorating costs across these various activities, is shown in Table 3.

Costs for tests here are simply those incurred by the school district for tests it manages of overall achievement. In addition,

Table 3

METRO DISTRICT: COSTS OF TESTING PER PUPIL TESTED BY TEST

TEST	TOTAL COSTS	COSTS PER PUPIL TESTED[1]
Continuum-based SES district test	$ 467	$ 1.60
Norm-referenced test	162	1.55
Math Proficiency	173.5	3.55
Writing Proficiency (Junior High)	101.5	2.13
Reading Proficiency (Junior High)	181.5	3.68
Math Proficiency	175.5	2.97
Writing Proficiency (Senior High)	102.5	1.79
Reading test (Senior high)	72.5	1.26
Total	$ 1436	Average $ 2.64

1 Numbers of pupils tested estimated using enrollments by grade level, plus estimates of test retakes for proficiency tests.

there are tests of language proficiency, supported by other funds, the annual state assessment activity for children in grades 1, 3, 6, and 12, and tests associated with college entrance, e.g., PSAT, or Advanced Placement examinations.

The above figures represent costs incurred outside of the school site. What happens when we include the costs of personnel time and other resources at the school level? We can provide a detailed answer for the elementary school we studied, a school which participated heavily in categorical programs.

Table 4 presents estimated aggregate costs of testing at a typical elementary school, estimates which when summed reach a figure well beyond $8,000,000 annually.

Table 4

ESTIMATES OF TOTAL METRO DISTRICT ELEMENTARY LEVEL REQUIRED TESTING COSTS AT CITYSIDE SCHOOL

Type of Costs		Total at Cityside	Per Pupil
District Office Cost			
2.64 per pupil X 830 pupils		$ 2,191	$ 2.64
Direct Costs to School			
Test Purchases		6,400	7.71
Indirect Costs for Schools (Personnel Time)			
	Hours/Year		
Administrators/Coordinators	95.6	1,650	1.99
Clerical/Secretarial	.51	5	--
Classroom Teachers	24.1	332	12.00
	X 30	X 30	
	723	9,964	
Instructional Specialists	90.4	1,580	1.90
Instructional Aides	220.3	1,322	.53
TOTAL COSTS FOR SCHOOL		23,548	28.36
Average Cost Per Classroom		785	

Assuming these costs are typical for all elementary schools, cost to district = $8,252,760.

Now these cost estimates, and their meaning, should be reviewed in the context of earlier concerns for test quality: That is, are the data any good at whatever cost? In addition, our ethnographic work in classrooms (Dorr-Bremme et al, 1983) revealed that many actual time costs continue to be underestimated by teachers as well as difficult to discern by observers. For example, when teachers know that a period of upcoming time will need to be be assigned to formal testing, they

wisely defer the beginning of a new unit of the curriculum and mark time until the testing period occurs so that the students may have more continuity in their instructional experience. This down-time cost of tests which should actually be included in teacher time estimates is not reflected in our data. Nor are the wide range of assessment activities developed by teachers included here. Other transparent costs include organization and follow-up of testing sequences, conference time with parents, etc.

We also attempted to look at other more direct opportunity costs. Clearly, even by the conservative estimates of time by teachers, 24 hours a year is devoted by a given teacher to required tests. This number takes on some significance if one remembers that somewhere around 175 hours is the total required allocation for reading instruction required by the state in which Metro District is located. Moreover, classroom observation studies (Rosenshine, 1980) have demonstrated that only about 40% of time allocated to particular instructional goals like reading and math actually consist of students spending time in direct or related instructional activities. The rest of the time is divided among procedural and management tasks, interruptions, and irrelevancies. Cutting formal testing time in half could add 20% to the amount of focused teacher time available for reading instruction.

Another way to look at the total cost is in terms of how much money is discretionary. Certainly, most educators are aware that the bulk of school district budgets is allotted proportionately to fixed

and inflation-sensitive costs of personnel. Over one million dollars could be reassigned to schools in the lowest quartile.

Last, we are not, at this time, in a position to develop the arguments around other test costs and benefits, e.g., anxieties and incentives, that undoubtedly influence the effectiveness of testing practices.

Additional costs not clearly estimated in this case derive from the allocation of student and teacher time for test retakes. More important, and missing as well, are the costs in student and school personnel time of the shadow-curriculum spawned by some proficiency tests. When students do not pass such tests on their first trial, many districts have remedial programs in which students are provided practice in skills assessed on the measures. This practice typically takes place outside of the normal curriculum in use for students generally. So rather than revising the mainline curriculum and instruction so that more students will be succesful the first time, a parallel, remedial curricular effort is mounted. The reasons for this decision are many, included in which is the need to show special program allocations directed specifically to bringing students with performance deficits up to par. What curriculum opportunities are lost by these students vary from place to place. Football is not the only or most significant casualty, and reduction of time spent in the humanities has been common.

Another potential cost is the effect of the selection of proficiency standards for student performance. The tradeoff here is obvious. If one chooses a standard, like 9th grade performance, as the

minimal competency level expected of all students, one can raise automatically the number of students who will pass at first or pass ultimately the tests. The unavoidable question is "What does the school do with 10th, 11th, and 12th grade if 9th grade proficiency is all that can be expected?" Raising the standard to average 12th grade performance would elevate public perception of school effectiveness, but would drastically reduce the number of students who pass. No satisfactory option seems to exist for the schools. In addition, a review of high school-leaving data suggests that school populations in urban settings may decline almost by half from the 9th grade (where proficiency is first measured) to high school graduation (Baker, 1982). So while the district can and does report proportion of students passing the proficiency examinations at rates in excess of 95%, a very much smaller pool of examinees contributes to this impressive figure.

Finally, in reflecting on the Metro example, it is important to remember that the testing here is principally employed as a measure of output, for externally driven accountability concerns rather than as an intervention intended to inform fully the design, revision, and implementation of a high quality curriculum program. What also should be remembered is that the Testing and Evaluation office at Metro is competently led and sensitively managed and that many school districts use up proportionately more dollars less efficiently.

Serious Test Use

If Metro represents a large bureaucratic effort to use tests as output measures, one still can and should raise the question of whether

any improvement occurs, either on the criterion measures themselves or noticeably on anything concrete as a consequence of the testing program. Simply, the question is, does testing have positive effects on school performance?

Part of the CSE Test Use study focused on the use of tests made by teachers, using both self-report survey data and observation and interview of classrooms and teachers. A principal finding was that left to their own preferences, most teachers do not put much faith in information derived from formal testing (Herman & Dorr-Bremme, 1983). The reasons are well known: test information is rarely available for teachers, even if they wanted to use it, on the schedule or in the form that allows its easy integration with instruction. Information is rarely diagnostic, that is, it often does not tell what is wrong and needs additional attention, but instead merely points out general areas of deficiency. Consequently, teachers are much more likely to use test information as corroboration for their own estimates of student need.

To conceive of a testing program that within acceptable cost boundaries allows a significant improvement over what teachers ordinarily do represents a long-term goal of research at CSE. In order to understand more fully the options and weaknesses of various approaches, consider two extant alternative ways of using tests principally as interventions to improve systems performance rather than merely as means to comply with general information and compliance requirements in accountability contexts. The first example derives from a political imperative to raise district test scores; the second is a more serious attempt to use testing to improve instruction and learning.

Political Improvement: Tests at Seaview District

Seaview's approach to testing was a more self-conscious attempt to use test performance to drive improvement in the schools. In the previous example, Metro District, because of its sheer size, complexity, demography, and transiency is relatively insulated from strong performance demands. Seaview District is altogether different. Composed of 14 schools, including two junior high schools and one senior high school, Seaview is a manageable size, with relatively little transiency and a predominantly Anglo student body. Unfortunately, however, the performance profiles of this school district on statewide assessment were always well below average, even when corrected for socio-economic status. The school superintendent had, at least, a political problem. He had to demonstrate that direct action in the form of increased school services had effects on subsequent student performance on a statewide measure. To assist the district, a team analyzed data patterns available for the 11 elementary schools and adopted as a goal the improvement of test profiles. This instrumental goal was chosen even though it was well understood that effects on real and important learning might be traded-off for better test scores. The team looked at the percentile rankings of the schools against various subscales on the measures and noticed, not surprisingly, that lowest percentiles were obtained on skills where raw scores were both high (around 80% correct) and bunched close together. Thus, the district was encouraged to emphasize those areas where real growth would be possible (having regression effect work for them) rather than to address those areas

where percentiles were especially low. In addition, subgroup analyses were conducted, and counter to common belief, the district was performing rather well when its poorest students were considered. In fact, the deficit in performance could be substantially allocated to sub-average performance by the higher SES students, those whom we would have expected to perform well. Thus, these children were selected as the principal target group. Test areas for emphasis were determined by reviewing the published test specifications and choosing areas that 1) had apparent transfer power, e.g., structural affixes showed up on both the reading and language examinations; and 2) those specifications that seemed to be amenable to direct instruction rather than insight, intelligence, or talent. A review of texts was conducted with the finding that there was little correspondence between instruction found in the district reading program and skills measured on the test. Consequently, coordinated practice exercises for tested areas were developed and the principals at each school agreed to monitor teachers' use of the exercises. In addition, some standard "test-wise" practices were put into effect, e.g., higher adult-to-student ratios during test administration, practice on timed tests in formats similar to those tested, gentle redirection of attention to the test itself when children's concentration waned. The costs for such efforts are indicated in Table 5 below.

Table 5

Seaview

COSTS OF TEST IMPROVEMENT EFFORT

Services of Team Consultation Test Analysis	$ 7,500	
Development of Practice Exercises	7,500	
Principals' time: 3 hrs. at 11 schools at $25/hr.	825	
District Administrator time: 8 hrs. at $30	240	
TOTAL	$16,065	
- by 11 schools	1,460	per school
- estimated cost per child (100 students at target grade level per school)	$14.60	

The costs presented in this table are exclusive of teachers' time, since their efforts did not represent significant departures from current practices. Obviously, had practice materials been developed by teachers, in-service training costs, plus the cost of released time for participating teachers would need to be computed. Thus, the $14 plus dollars represent an additional cost, over and above standard district testing costs for improving performance.

To reduce excrutiating suspense, if any, the results were that all schools in the district improved performance on the next statewide test administration and were found to be within, rather than below, the confidence bands of achievement expected for similar schools.

Very Serious Test Use: Student Achievement Model*

The student achievement model describes a system developed over the past dozen years in a small California school district as part of its curricular emphasis on individualized instruction. In this district, teachers use the results of semi-annual, criterion-referenced achievement testing in order to place students in classes, to group youngsters within classes for instructional purposes, to assess the effectiveness of their curricular strategies, to prescribe remedial activities when needed, and to provide a basis of communication with parents. While all of these purposes are commonly attached to school district assessment practices of one type of another, the study district illustrating this model incorporates these objectives into a tightly linked "system." Not only are these purposes served by a single battery of tests, appropriately geared to grade levels, but also the district's core instructional continua in reading, mathematics, and language skills have been developed in tandem with the tests by the district's teachers and staff. The instructional program and the assessment instruments are thus intentionally matched, and the information generated by the assessments is viewed commonly by district personnel as both relevant and salutary for instructional planning and improvement.

Pupils are tested in the fall and winter of each school year. Results of tests, scored and elaborately organized through district

*Exerpted from: Catterall, J. The cost of instructional information systems: Results from two studies. CSE report no. 208. Los Angeles, CA: Center for the Study of Evaluation, University of California, 1983.

data processing services, are available to teachers within a week. Learning specialists at each of the district's schools assist in test administration and interpretation of results. Principals use the test results as the primary basis of fall and spring planning sessions with individual teachers -- a critical component of the district's instructional leadership activity.

The student achievement model is now a continuing, stable, and dominant fact of the district's instructional life. Both daily instructional activities and incidental assessments of pupil progress are directly geared to the scope and sequence of topics outlined in the continua. Teachers and administrators universally report the centrality of this "system" to their efforts. System costs at the district level are depicted in Table 6.

Costs at the school site level are greater than those at the central office level, primarily because of the significant amount of time spent by teachers and principals in planning instruction on the basis of system reports. Each principal spends a full week twice per year in one-to-one consultations with teachers to assist in instructional management. For each principal, this activity contributes to a total of more than a ten percent allocation of time on a yearly basis to the system. For each teacher, this planning activity occupies about six hours per year. In addition, a learning specialist at each school site devotes one day every other week to system activities. These school site-level costs, shown in Table 7, amount to a total of about $15.00 per pupil over the year.

Table 6

STUDENT ACHIEVEMENT MODEL: CENTRAL DISTRICT COSTS

	Type of cost	Cost Estimate
A.	Personnel	
	° Evaluation Coordinator (17.5% FTE @ $34,000	$ 5,960
	° Instructional Materials Coordinator (5% FTE @ $30,000)	$ 1,500
	° Assistant Superintendent (5% FTE @ $40,000)	$ 2,000
	° Clerical Support (37.5% FTE @ $18,000)	$ 6,750
	Total Personnel Costs	$16,210
B.	Equipment and Materials	
	° Computer (17.5% devoted to CRT, annualized cost of ($10,000)	$ 1,750
	° Paper and Materials	
	- Answer sheets	$ 750
	- Photo copying	$ 750
	- Printing	$ 2,500
	Total Equipment and Materials Costs	$ 5,750
C.	Total Central District Costs	$21,960
D.	Per Pupil Cost	$ 4.22

Table 7

MODEL: SITE LEVEL COSTS

Type of cost	Cost Estimate
A. Principal (2 weeks plus 1/4 to 1/2 day per week ongoing = 12% FTE @ $30,000)	$ 3,571.00
B. Learning Specialist (10% FTE @ $28,000)	$ 2,800.00
C. Media Specalist (2% FTE @ $25,000)	$ 500.00
D. Teachers (6 hrs. @ $17 for each of 22)	$ 2,244.00
E. Total Cost	$ 9,115.00
F. Per Pupil Cost	$ 15.19

The remaining costs of the system are displayed in Table 8. In addition to spending 5 to 10 hours per semester in administration of the tests, teachers spend about 5 hours in preparation and grouping youngsters for testing, and some teachers receive brief inservice sessions related to the testing program. In addition, some items on the tests are teacher- or aide-scored, and the values of these time allocations are shown in the table. The time pupils spend taking tests and the time of parent volunteers have been recorded in the table, but no dollar approximations have been made. Testing costs identified amount overall to a little more than $17.00 per pupil.

Table 8

MODEL: SITE LEVEL TESTING COSTS (PER 30 PUPILS)

	Type of cost	Cost Estimate
A.	Pre-Test Activities	
	° Teacher planning: 5 hrs. (@ $17 per hr.)	$ 85.00
	° Teacher inservice: 1/2 hr. (not all teachers each year)	$ 8.50
B.	Test Administration	
	° Teacher: 15 hrs. per year average (5 to 10 hrs. per semester)	$ 255.00
	° Pupils: 15 hrs. per year	$ n.b.*
C.	Scoring and Analysis	
	° Teacher-scored items: 8 hrs. per year average	$ 136.00
	° Aide: 4 hrs. @ $10 per hr.	$ 40.00
	° Parent Volunteers: 4 days	$ n.b.*
	Total Testing Costs (30 pupils	$ 524.50
	Testing Costs Per Pupil	$ 17.48
D.	Total District Testing Costs (3800 pupils)	$60,600.00

*n.b. = Non-Budget Item

The total costs are summarized in Table 9. Central office, school site, and pupil testing costs total about $34.00 per pupil in the district. To this figure we might add a factor representing the value of pupil time involved for testing (about 15 hours per year) to achieve an overall picture of resources supporting the testing system.

Table 9

MODEL: TOTAL SYSTEM COSTS

	Type of cost		Cost Estimate
A.	Central Costs		
	° Personnel		$ 16,210.00
	° Equipment and Materials		$ 5,750.00
		total	$ 21,960.00
B.	School Site Level (non-testing)		
	° Coordination and development		$ 63,805.00
C.	School Site Testing (524.50 per 30 pupils)		$ 60,600.00
		Total	$146,365.00
D.	Total Costs Per Pupil		$ 34.00

It should be noted that these cost figures do not include start-up costs of test development, revision, administration, or planning for implementation. Secondly, the test development was done in-house by teachers, a fact which invariably and unfortunately raises the question of quality control. One should also be reminded that these costs, just like those in the Seaview district, represent only a portion of the entire testing program.

In order to get a sense of what start-up costs might be, CSE work with an Eastern metropolitan school district, focusing only on writing and reading at the elementary level, cost $450,000 for special services over a three year period. These figures do not include a substantial amount of untracked administrative time or the cost of teacher release time.

Summary: What Now?

The lessons of this discussion are simple. First, testing costs money, and in times where discretionary dollars are constrained,

even relatively small figures need review and reflection. Second, the two more serious efforts, where testing was conceived as an intervention as well as an output measure, cost substantially more than one might imagine, with start up costs easily tripling the Metro estimates.

Last, one cannot evaluate costs without some notion of benefits. There is some question as to what benefits accrue from testing in general or from the testing interventions described here. From an accountability perspective, concerns raised earlier about the quality of available tests and their ability to serve as valid indicators of school outcomes must temper estimates of benefit. Not only are the instruments imperfect, in some cases grossly so, but they address only a relatively narrow band of skills and knowledge which can be conventionally and conveniently assessed. Ignored in general, for example, are higher level thinking, problem solving, and communication skills. The benefit from such measures, then, is some information about students' performance on a constrained set of indicators.

The breadth of focus issue also impinges on benefits derived from testing interventions designed to facilitate student learning. One wonders whether such interventions in practice actually target essential components of complex student performance or whether they simply emphasize skills and competencies which are easily both specified and assessed. While there is some evidence that district testing programs can help increase student achievement on like objectives (LeMahieu, 1983), there is also some question regarding the transfer value and cognitive level of such performance and some

attendant danger that instruction in higher level skills is being replaced by practice in rote learning.

Our conclusions have broad implications for current debates about excellence in education. First, we must question the degree to which the costs of extensive assessments in the schools are incurred at the expense of other activities which might lead to advances in achievement. Both budget allocations and pupil and teacher time are critical resources which may be assigned in many ways to promote learning. Second we must also question the degree to which future positive trends in large scale assessment results -- if any should follow as we enact our designs for excellence -- are truly indicators of quality learning. We have implied here that alternative interpretations, political and professional glee notwithstanding, might be more justified.

References

Baker, E.L. Lassen high school. Report to the Carnegie Foundation. Los Angeles, California: Center for the Study of Evaluation, University of California, 1982.

Baker, E.L. Evaluating educational quality: A rational design. Invited address to the Center for Educational Policy and Management, University of Oregon, 1983.

Burry, J., Catteral, J., Choppin, B., Dorr-Bremme, D. Testing in the nation's schools and districts: How much? What kinds? To what ends? At what costs? CSE Report #194. Los Angeles, CA: Center for the Study of Evaluation, University of California, 1982.

Dorr-Bremme, D., Burry, J., Catterall, J., Cabello, B., & Daniels, L. The costs of testing in American public schools. CSE Report # 198. Los Angeles, CA: Center for the Study of Evaluation, University of California, 1983.

Floden, R.E., Porter, A.C., Schmidt, W.H., & Freeman, D.J. Don't they all measure the same thing? Consequences of standardized test selection. In E.L. Baker & E.S. Quellmalz (Eds.), Educational testing and evaluation: Design, analysis, and policy. Beverly Hills, CA: Sage Publications, 1980.

Herman, J.H., & Cabello, B. A comparison of CAP and a major reading text at grade three. Los Angeles, CA: Center for the Study of Evaluation, University of California, 1982.

Herman, J.H., & Dorr-Bremme, D.W. Uses of testing in the schools: A national profile. In W. Hathaway (Ed.), Testing in the schools. New Directions for Testing and Measurement. San Francisco: Jossey-Bass, Inc., 1983.

LeMahieu, P.G. A study of the effects of a program of student achievement monitoring through testing. Doctoral dissertation, University of Pittsburgh, Pittsburgh, PA. University Microfilms, Ann Arbor, MI, 1983.

Rosenshine, B. How time is spent in elementary classrooms. In C. Denham A. & Lieberman, (Eds.), Time to learn. Washington, D.C.: National Institute of Education, 1980.

Walker, D.F., & Schaffarzick, J. Comparing curriculua. Review of Educational Research, 1974, 44, 83-111.