CAN WE FAIRLY MEASURE THE QUALITY OF EDUCATION?

CSE Technical Report 290

Eva L. Baker

Center for Research on Evaluation,
Standards, and Student Testing
UCLA Center for the Study of Evaluation

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.

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

Introduction[1]

American political attention has turned with increasing intensity to the matter of educational quality. From the reports of commissions and panels to debates by presidential candidates, the focus on students, teachers, and schools grows sharper every day. At the center of concern is a deceptively simple question: How well do our schools prepare our students?

It doesn't matter if the language emphasizes excellence, subject matter understanding, productivity, or competitiveness, the meaning of the debate is clear: Can we describe, judge and improve the effectiveness of public schools?

Over the years, significant investments have been made in trying to answer these questions. Standardized achievement tests, educational program evaluations, teacher testing, and minimum competency tests for students all are thought to provide useful information to help make judgments about the effects of educational services on students. Many of these options have roots in the mid-sixties enactment of federal legislation to assist educationally disadvantaged students. This new legislation required that the federal government evaluate the effects of its efforts to provide compensatory resources for students. The legislation was directly responsible for the rapid development and growth of the evaluation field and for many scientific developments in the measurement of human performance. Through the ensuing decades, one or another particular version of evaluation or measurement was selected as the new solution for understanding school effectiveness, the options coming, it seemed, in overlapping waves. Remember? Different solutions included setting objectives and measuring student performance, local standardized student testing, program evaluation, Scholastic Aptitude Examination (SAT) score decline, state minimum competency examinations, teacher testing, state assessment, and "The Wall Chart," a national comparison of educational systems. None of these approaches were found to be wholly satisfactory, but, after the initial blaze of interest died down, none were retired either. Instead, our attempts to understand educational quality have resulted in an increasing set of measures and approaches designed to shed some light on the issue. But do they? Imagine that we could start over, fresh and unsullied by our prior measurement experience. What would be fair measures of the effectiveness of our educational programs?

To answer this question, we first must decide what level of information we want. Making a judgment about all of American education and assessing the effectiveness of First Street School in your hometown require different levels of information. In the first case, we would look for common features of schools and curricula to base our judgment. When looking at a particular school, however, we can be much more attentive to the community characteristics, the kinds of students attending the school, the particular goals of the school, and other special conditions. In both cases, however, we simply want to know the following:

What are the students learning?
How well do the teachers teach?
What is the quality of our schools?

The public seems equally interested in the concrete accomplishments of local schools and the general descriptions of the educational system at large.

Educators want answers to these questions. These answers should not simply describe the state of performance for students, teachers, and school administrators, but should ideally permit us to devise actions to make things better. We want information for more than curiosity's sake; we want it to help us improve education. This desire to

---

[1]A version of this paper will appear as an article in *NEA Today: Issues '88, 6*(6), pp 9-14.

face and fix what's wrong requires that the information we collect gives us more than categorical "good" or "poor" labels. We need enough detail to guide our policies and practices.

With this discussion as preface, let's consider in turn questions of effectiveness that involve students, teachers, and schools.

Student Learning

Student learning has been traditionally measured by achievement tests. For public accountability purposes, teacher-made tests have never been regarded as sufficient. Rather, because accountability implies some sort of comparison, tests that have standard content and rather general applicability have been used. Without rehashing two decades of concerns about standardized testing, a few issues remain salient:

1. Standardized tests allow comparisons among schools and regions. They may, however, be somewhat insensitive to curricular and instructional variations. Because they are prepared to be of widest utility, standardized tests may omit areas of particular emphasis for particular schools. These tests provide information on only a narrow slice of school activities.

2. Standardized tests most often ask children to answer questions given in multiple-choice format. I believe this format greatly underestimates student performance.

3. Because of technical reasons used in test statistics, very small absolute differences (for instance, one test item) might mean an improvement of a "grade level" or so. Making inferences about educational quality based on these differences is a shaky proposition.

Test performance still is, in that unfortunate phrase, the bottom line for many who would assess the effectiveness of the schools. At this time, standardized tests are regarded by many policymakers as credible and objective. Achievement testing will not go away, and for good reason. Students and, by implication, the schools to which they go must be held accountable for teaching students and for attempting to measure what they have learned. Standardized tests are thought by many to be the best approach we have.

But these tests can be greatly improved. At the Center for Research on Evaluation, Standards, and Student Testing (CRESST), sponsored by the U.S. Office of Educational Research and Improvement, we are in the midst of a five year research program to improve the quality of testing for use in the schools.

The precepts of our program, and the way we believe testing ought to be improved, fix on a small set of critical issues. In one way or another, our attention focuses on validity, or the quality of the information the test provides us and the degree to which we can believe it.

Validity

Validity of achievement measures has a number of components (see Baker & Herman, 1983, for a fuller discussion). One critical component is the degree to which the way performance is measured matches the mode in which learning best occurs. With the advances in cognitive science, we believe we can design measures that more

productively represent the richness of learning. For example, we are interested in assuring that in mathematics, science, and history, students be given different ways to demonstrate their competence, perhaps in multiple choice tests, perhaps in other paper and pencil formats, perhaps using computer dynamic displays, perhaps in writing. Many current testing formats developed out of convenience for the administration and scoring processes rather than because they were the best ways to assess complex human understanding. One attribute of tests is that they often force students to give the first, quick response, rather than a thoughtful, reasoned answer. The balance between conserving the time spent on testing and providing enough opportunity for adequate thought is still unsettled. Perhaps a more diverse menu of testing approaches will increase the overall validity of our measures, and allow testing approaches to match better student propensities.

A second validity concern relates to the content or subject matter of what is to be tested. One of the sadder outcomes of the behavioral objective movement and of inquiry approaches of the early seventies was the attention paid to process *at the expense* of the content to which these processes applied. We have seen the pendulum swing widely on this issue during the last two decades. Given the popularity of books like *Cultural Literacy* (Hirsch, 1987) and the scandalous blanks and misunderstandings in our students' knowledge, we are again on the verge of another swing towards content. It's tempting to devise tests that can pinpoint such content errors. This time, however, we want to assure that we go well beyond identification or recognition of specific facts and concepts. We intend to integrate measurement approaches that wed content with sophisticated approaches to demonstrating understanding, such as complex essays. We at UCLA are developing the technology to score such essays reliably and relatively inexpensively.

Third, we are interested in measures that can be related directly to instructional options. We should be measuring performance that schools can affect. This means that, where possible, we should be collecting information about teaching practices, student familiarity with content, and so on, at the same time that we measure student performance.

Fourth, our measures must be valid when individual and group differences are considered. Whether a test is fair is a psychological as well as an empirical issue. We particularly want to assure that our measures validly assess strengths and weaknesses of our pluralistic student body in a way that contributes to their motivation to continue learning.

Quality of Interpretation

Even when student achievement is measured validly, the way such findings are interpreted makes a difference. Interpretation involves relating findings to other similar measures of performance, comparing findings to the performance of other similar groups of students or schools, analyzing findings in the light of previous performance to see the development of trends over time, or looking at performance in terms of some predetermined standard. Comparison to other student groups is the most common interpretation strategy. This comparison is the basis of "norms," or averages provided for many nationally standardized tests. In some state assessments, comparisons for student performance are provided by looking at the performance of students in schools of similar size and community location. More recently, the federal government has reported the comparison of student performance on the SAT state by state, a specific approach to be discussed later.

A central issue of interpretation is what is being compared. Are tests of individual students used to make comparisons among schools? What other information needs to be collected if such use of information is to be sensible?

The first question for these sorts of comparisons is: "Is the comparison fair?" One shouldn't compare a small, stable suburban school with a central city school that has a high mobility rate. Given the increasing diversity of our students, comparisons now must involve issues such as language in the home and length of time in the school in addition to the more usual socioeconomic measures.

Other options have been the international comparisons, where we look at U.S. students in comparison to those in other countries. While such comparisons might be useful in setting goals for our students, the inference remains that we should adopt practices embedded in other cultures or in other constitutional, and more centralized, arrangements for education policymaking. Such an inference is probably unwarranted.

Moreover, the bane of most normative comparisons is that half of the group is always below average, a status unacceptable to most educational policymakers. No one yet has figured out how all students can perform "above the average."

To sum up, what should we want in student achievement measurement?

1. More than one measure of the same phenomenon, such as reading comprehension (to allow for corroboration from different sources), but with no expectation that all students need to take multiple measures.

2. More than one kind of testing format, such as multiple choice and written answers.

3. Tests that give students adequate time to perform serious cognitive tasks.

4. Tests that measure both the content (what) and the process (how) that students use to solve complex problems of understanding.

5. Tests that can be analyzed to guide instructional planning.

6. Test results that are understandable, timely, and usable by teachers for instruction and planning (see Herman & Dorr-Bremme, 1986, for a report of teachers' test use.

7. Reports of test results that are fair to students, teachers, and schools.

Quality of Teaching

A second enduring concern in education is the quality of teaching. This interest is obvious; when we think of schools we think of teachers. Given the instructional and economical dominance of teachers in schooling, it's natural to want to judge effectiveness of educational investments in part by looking at teaching. The problems begin when one tries to operationalize the measurement of the quality of teaching and confuses it with the "quality" of teachers. Just as in the student achievement area, the principal trouble spot in quality of teaching is validity. There is little real agreement on what good teaching is. When good results occur, we can attempt to infer which teaching practices were responsible. A general application of principles such as providing students with opportunity to learn, clear task directions, and feedback undoubtedly apply on the average. Our problem is that we are often not interested in teaching on the average, but are particularly interested in a particular teacher's competency, perhaps for merit pay or other forms of advancement. When the individual teacher is our focus, we must take special care to allow adequate flexibility in

pedagogical style, since for various topics, objectives, grade levels, personalities, settings, and student groups, no "best" pedagogical approach has been identified. With support of the Carnegie Foundation, new approaches to the assessment of teaching competencies are under development. Although designed to permit special certification of teachers rather than the assessment of educational effects, their efforts may have some positive influence on the measurement of teaching capability.

Teacher Testing

Because teaching quality has been hard to measure, many have supported the measurement of prerequisites that good teachers are presumed to need. Such prerequisites include mastery of subject matter, mastery of basic knowledge about teaching, student development and learning, and mastery of basic skills. Tests have been devised to assess teachers in many of these areas, some with associated sanctions. Without disputing the right of the state or school district to set standards of this sort, conflicts have developed on a number of points. Rudner (in Office of Educational Research and Improvement, 1987b) points out that the standards for many of these tests have been set very low. Lorrie Shepard, in a case study of the Texas Teacher Test (1987), describes how it might be possible to pass the test by being testwise rather than being skilled in the area the test was assessing. Ellwein and Glass (1986) infer from their case study that teacher testing is mostly symbolic and has very little to do with actually identifying deficiencies and improving instruction. Involved in many of the analyses of teacher testing is the question of when it should occur (pre-service? pre-teacher education program?) and to whom the sanctions should apply (the teacher? the degree-granting institution? the teacher training institution?).

Student Achievement as a Measure of Teaching

Using student achievement as a way to estimate teaching effectiveness is another approach. It seems like a reasonable tactic; after all, teachers ought to help students learn. Clearly subject to the validity concerns about student testing listed above, the use of such measures to assess teachers unfortunately adds new complexity. Minimally, these comparisons may necessitate complex tracking of students who enter particular teachers' classes. Statistically equating students with different entry competencies is sure to be an unsatisfactory way to compare teachers' relative merit in promoting achievement. On the one hand, it's harder to teach students who have inadequate backgrounds. Alternatively, it's also difficult (because of artifacts of tests) to show real improvement when the student group comes in with a very strong achievement level. In either case, the achievement tests will probably misrepresent the nature of the teacher's effort. Thankfully, recent assessment systems for teachers are attempting to represent more broadly the nature of teachers' efforts.

Educational Quality of Schools

Who wants to know? The desire to find out how schools are doing is clearly legitimate, and educators, policymakers, and researchers continue to propose alternative sources of information. One of the problems we face is providing the right information to the right people. Congressional policymakers want to know whether the schools are working (Congress of the United States Congressional Budget Office, 1986). At different times, their concern may be focused on the quality of what is learned (as in the post-Sputnik period) or who is learning (when equity concerns are central on the educational agenda). Their needs are to assess the impact of resources they have invested and to target continuing or new needs. They need relatively unambiguous, clear information. To even a greater extent, state level policymakers are concerned with the effects of specific policies related to financing, curriculum, and certification (i.e., their efforts to reform schools in their states). Local school boards and their administrations have needs for information related to the quality of their policy

implementation and the progress toward discretionary goals, given the particular characteristics of their community. Each set of policymakers has differential need for detail and different opportunity to influence the reality of classroom practice. The hodgepodge of conflicting information from local, state, and national evaluations doesn't make evaluation of educational effectiveness any easier. Some new approaches may offer some relief.

Comparisons State by State

An approach under consideration by the federal government is to transform the measurement practices of the National Assessment of Educational Progress (NAEP) so that state-by-state comparisons may be possible. NAEP has been administering measures periodically to U.S. students in reading comprehension, writing, and mathematics on a regular basis. At the present time, the administration of these measures allows for interpretation by broad geographical region, rather than for each state. The proposal calls for administering these measures so that a representative sample of each state would be tested and described in NAEP reports. The proposal also expands the number of subject matters assessed. If accepted, this approach could focus the evaluation of schooling on the NAEP achievement measures. Is this a good thing? There is a clear division of opinion. Let me review some of the arguments on behalf of and against this approach. On the positive side:

1. A common basis for understanding student achievement would be systematically available.

2. The quality of measures would continue to improve because of the salience of the measures.

3. States could use such information for their own policy assessment to check their progress.

4. Interpretation for policy purposes would be simplified.

5. States would be able to compare themselves to subsets of other similar states.

On the negative side, critics content that:

1. NAEP may turn into a national achievement test, and a national curriculum may follow.

2. NAEP will not be sufficiently responsive to local or regional differences in curricula, students, or economic factors to permit legitimate comparisons.

3. NAEP will drive out state and local tests, which are more responsive to local curricula.

4. The pressure for school district comparisons will follow state comparisons.

5. Because NAEP's strength will be comparisons over time, the pressure to keep NAEP measures the same will inhibit new goals for the curriculum and new approaches to measurement.

6. A single set of measures can be wrong. Given the state of understanding of achievement measurement, investing in different assessment approaches is the most prudent way to collect policy relevant information.

For each of these points, both positive and negative, there are counter-arguments, and counter-counter arguments. If the problem were simple, it would already be solved. The attractiveness of a clearly understood, single set of measures for American education is strong, even when the validity of the measures for assessing local and state educational policies is questioned. The state-by-state NAEP approach needs to be understood as an attempt to catch hold of what our schools are doing.

Quality Indicators

Another tack is the quality indicators movement (Office of Educational Research and Improvement, 1987a). The goal of this effort is to identify and systematically collect information that can give a picture of the overall quality of American education, not simply limited to achievement testing. Work in this area has been conducted by The RAND Corporation, the Center for Policy Research in Education at Rutgers, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA, and by numerous other institutions and scholars. Part of its impetus comes from the realm of economic indicators, where seemingly simple numbers like the Gross National Product, unemployment figures, and the Dow Jones average efficiently communicate the economic health of the country. The Center for Education Statistics (a division of OERI), under the leadership of Emerson Elliott, has been working on indicators of educational quality. These indicators include figures such as dropout rates, per capita student funding, student-teacher class ratios, enrollment figures, and the like. Problems encountered with this approach include the vastly different reporting approaches taken for something as understandable as student dropout. Different districts and states count dropouts at different intervals, for different ages or grades, use different base rates, track student mobility differently, and so on. Getting everyone to agree on a single reporting approach, even for an "easily understood" concept like student dropout, is a Herculean task.

Outcomes like achievement test scores, college admission rates, or dropout figures represent the easy part of indicators. Quality indicators should also take into account input variables and measures of process.

Imagine one wanted a "quality indicator" related to some intermediate process, such as student coursework. In fact, UCLA and The RAND Corporation are collaborating on the development of such indicators. We need to consider how to determine "quality" in a valid and comprehensive way, how to collect such information accurately and comfortably in schools, and how to report such findings so that the effects of educational reform can be tracked. If we (or others) can solve such a problem, educational achievement tests can be relieved of the perhaps excessive burden they carry as measures of the effects of different policies. Making changes, such as adding coursework requirements, strengthening the content of the curriculum in a particular area, or requiring textbooks to exhibit certain content standards, are all hypotheses that policymakers make about what will help schools. Indicators of the extent to which these policies are used is a first step; studying the relationship of the level of their use and resultant levels of student achievement is a second critical link. Yet, the indicator movement must be cautious about identifying a single magic index or number to stand for complex educational processes. As Leigh Burstein of UCLA points out, the context in which such data are reported, understood, and interpreted, is central to the success of this effort (Baker, 1987).

Summary

The search for approaches to assess schools, their teachers, and students will continue. This discussion has touched lightly on a number of complex issues. Controversies also will continue, and we can be sure that almost any decision will be rethought sometime in the future. Our interest in the research community is to keep a few issues in front of the public and the decisionmakers in this area.

First, we believe that the validity of any measure or indicator should be paramount, whether it is a measure of outcomes, like student achievement, of input, like teacher knowledge, or of processes, like student coursework. These measures should be designed to allow multiple or flexible ways to demonstrate success for different students. These measures should help us to pinpoint and fix weaknesses in policy and practice. Finally, these measures first must serve the interests of students and improve their schools. We must overcome our habit of preparing measures for the convenience of test developers, administrators, legislators, or even teachers. Rather, we need to consider the impact of our approaches to assessing educational effectiveness on our current and future students.

References

Baker, E.L. (1987). [Interview with Leigh Burstein, Professor, UCLA Graduate School of Education, and Study Director, Center for Research on Evaluation, Standards and Student Testing, UCLA Center for the Study of Evaluation.]

Baker, E.L., & Herman, J.L. (1983). *Task structure design: Beyond linkage* (CSE Report No. 199). Los Angeles: UCLA Center for the Study of Evaluation.

Congress of the United States Congressional Budget Office (1986). *Trends in educational achievement* . Washington, DC: Author.

Ellwien, M.C., & Glass, G.V. (1986). *Case studies on education standards.* Los Angeles: UCLA Center for the Study of Evaluation.

Herman, J.L., & Baker, E.L. (1985). Educational evaluation: Emergent needs for research. *Evaluation Comment, 7*(2). Los Angeles: UCLA Center for the Study of Evaluation.

Dorr-Bremme, D.W., & Herman, J.L. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph Series in Evaluation No. 11). Los Angeles: UCLA Center for the Study of Evaluation.

Hirsch, E.D. (1987). *Cultural literacy.* Boston: Mifflin.

Oakes, J. (1986). *Educational indicators: A guide for policymakers* (CPRE Occasional Paper, No. OPE-01). Los Angeles: The RAND Corporation.

Office of Educational Research and Improvement, U.S. Department of Education (1987a). *Elementary and secondary education indicators in brief.* Washington, DC: Author.

Office of Educational Research and Improvement, U.S. Department of Education (1987b). *What's happening in teacher testing: An analysis of state teacher testing practices.* Washington, DC: Author.

Shepard, L.A., Kreitzer, A.E., & Graue, M.E. (1987). *A case study of the Texas Teacher Test: Technical report.* Los Angeles: UCLA Center for the Study of Evaluation.