Education Policy Analysis Archives

Volume 3 Number 6 **March 3, 1995** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.

Editor: Gene V Glass,Glass@ASU.EDU. College of Education,
Arizona State University,Tempe AZ 85287-2411

Educational Assessment Reassessed: The Usefulness of Standardized and Alternative Measures of Student Achievement as Indicators for the Assessment of Educational Outcomes

William L. Sanders
Sandra P. Horn
University of Tennessee
sphorn@sacam.oren.ortn.edu

Abstract:

For decades, the assessment of educational entities--school systems, individual schools, and teachers--has evoked strong and sometimes violent emotions from the educational community, the general public, and their legislative representatives. In spite of attempts to codify standards for the evaluation of these entities, assessment experts remain denominationalized--often religiously so. Methods of assessment based on the use of standardized tests have come under intense fire in recent years with some critics going so far as to call for their complete elimination. Those who advocate alternative methods of assessment have become increasingly outspoken in establishing exclusive rights to the legitimate assessment paradigm. However, some of the most respected advocates of alternative assessment have taken a more moderate view, warning against an "either-or" mentality (Brandt, 1992, p. 35). Reflecting this more moderate perspective, this paper strongly advocates the use of multiple indicators of student learning, including those provided by standardized tests.

The Debate

No responsible person claims that any form of assessment can appraise the totality of a student's school experience or even the entirety of the learning that is a part of that experience. However, it is possible to develop indicators to measure learning along important dimensions, closely related to the curriculum, both in standardized assessment instruments and in alternative forms of assessment. The real issue is not whether standardized assessment or alternative

assessment is the better model in every case for the evaluation of educational outcomes. Rather, the issue is choosing the most appropriate indicator variables for the specific purpose at hand, whatever that may be.

Non-standardized assessment is the traditional form of assessment within classrooms. Teachers construct questions, evaluate student responses, assign and check homework, monitor projects, and informally assess student progress hundreds of times a day. These assessments may be accurate or they may be faulty, depending upon the teacher's skill as a judge of various indicators and their applicability to the question at hand. However, unrelated factors can affect this type of assessment. For instance, research has shown that good handwriting can influence grading on essays (Feinberg, 1990, p. 15). Student behavior and appearance can likewise prejudice assessment. On the other hand, careful attention to the information provided by student actions and products can lead to a deep and revealing understanding of students' comprehension and skill attainment.

Standardized tests, whether the ubiquitous multiple choice test or other forms of standardized assessment, vary in their ability to fairly assess student knowledge, just as teacher assessments do. But current construction practices insure that standardized tests are subjected to rigorous validation criteria, reliability testing, and standardization procedures. The tests are open to review by experts and to criticism by anyone with credible grounds from which to argue. In the past, standardized tests have proved useful in comparing, generalizing, and indicating levels of attainment based on set standards. In current practice, they serve many additional functions including the assessment of higher-order reasoning skills and academic growth over time.

Non-standardized alternative assessment (referred to, henceforth, simply as "alternative assessment"), has yet to demonstrate the ability to provide generalizable information for comparison purposes over time on a large-scale basis without proving more costly in time and resources than standardized testing and without itself falling prey to the "teaching to the test" syndrome so often cited as a major deleterious result of standardized testing. The strengths of alternative assessment lie in its ability to individualize assessment, to mimic good teaching practices, and to involve teachers more deeply in the assessment process. Currently, attempts are being made to improve the generalizability and reliability of alternative assessments in order to use them for the evaluation of school and school system efficacy.

The Negative Perception of Standardized Measurement

Problems With Roots in the Past

Ignoring the strident voices of those who deny that effective educational practices can be examined at all, in hopes that it will be possible to reach a more useful conclusion, it must be acknowledged that the assumptions underpinning many of the evaluation schemes of the 70's and 80's were spurious. Of the doubtful assumptions, the most obvious were these two: (1) there is a right way to teach and (2) good educational practice can be identified independent of any demonstrated relationship to student learning. Although the public (erroneously) bought into the first assumption, they were never convinced of the second, and rightfully so.

The major indicator that distinguishes effective from ineffective educational practice is whether students learn that which is purportedly taught. However, student achievement data have rarely been incorporated in models for teacher and school evaluation, primarily because of the difficulty of delineating teacher and school effects on student learning from demographic effects--those effects that are embodied in the individual child-life of each student, independent of formal education.

Informally, however, the public, through the media, assesses the success or failure of the schools by how well students perform on standardized tests. The media uses test scores to

compare one educational entity to another, and, in a comparison, someone always loses. The teacher-made assessment which forms the bulk of student evaluation does not figure at all into the media's analysis of school success, and educators are unconvinced that group-administered tests provide an accurate indication of what a student knows or has learned under their tutelage. Therefore, educators have come to view standardized tests as the root of the public's disenchantment with the schools and the prime reason for the denigration of the teaching profession. It's no wonder there has been such an outcry against their use. Nevertheless, before we throw out the baby with the bath water, it would be well to look more deeply into the tub.

Recent Developments in the Use of Standardized Data

The negative perception of standardized testing as a measure of educational efficacy is, as stated above, partially a result of the way the results have been interpreted by the media and the general public. This problem is only indirectly associated with standardized testing. Possibly, with time and effort, the media can be educated as to the proper use and interpretation of the data derived from test scores. Regardless, this aspect of the debate must be relegated to the realm of societal problems rather than problems with the tests themselves. However, the perception that standardized scores can tell us very little about anything other than the status of a given student at a given time in regard only to the items on a specific test may derive directly from the historical uses of standardized data. This criticism is understandable in light of the past, but there have been a great many changes in recent years in the development of tests, the analysis of the data they render, and the uses to which they are put.

Modern standardized tests reflect a response to past criticism. Item response theory informs the construction of tests that are equivalent but non-redundant, thereby addressing the question of the score inflation that results from administering the same test, year after year. Test makers have also proved responsive to the need to assess higher-order thinking skills, so tests that contain items requiring application, analysis, synthesis, and evaluation are now readily available.

Standardized test data is ubiquitous. It is readily available, cheap, and abundant. Most often, the scale scores, percentiles, and stanines provided by those who score the tests are duly recorded by the receiving school or system, precisely as received. No further analysis is attempted. The usefulness of the data is severely limited, and so it is used for very little other than placement and, occasionally, a crude form of program evaluation that is, as critics have rightfully stated, biased by extraneous factors such as socio-economic level, past achievement, and percent of minority students.

The reasons for the misuse and lack of use of standardized test data are readily discernible. First, performing the operations necessary to analyze test data beyond the provided measures required an inordinate amount of time and expertise prior to the recent advent of powerful and inexpensive computers. It simply wasn't feasible for schools or systems or even states to proceed beyond the basic measures provided to them by the normal scoring algorithms. Merely constructing and maintaining a data base of schools, teachers, and students used to be an impossible task for all but the smallest systems. Constructing such a data base for a state was altogether unthinkable. Today, none of these obstacles are insurmountable.

Second, there were enormous statistical problems involved in the use of test scores for evaluative purposes. Among these were the regression to the mean and the problem of missing data. There was also the problem of delineating educational influences from the influences of extraneous factors. Now, there are statistical models that can deal effectively with all of these difficulties. One of these is briefly described later in this paper.

The refinement of test construction techniques, the widespread use of powerful computer technologies, and the application of sophisticated statistical methodologies in education is

producing a revolution in the use of standardized test data in educational assessment, but the misuse and under-utilization of test data in the past haunts the present. Standardized test scores, when subjected to appropriate methodologies and utilized for appropriate purposes, provide rich data for educational assessment. It would be lamentable if the perspective these data can afford were obscured by arguments no longer valid.

The Importance of Appropriate Indicators

The answer as to what is happening in any situation is dependent upon the questions asked--not just what is asked but also how it is asked. The answer will be more or less precise depending upon the means used to gather data and the extent of the data gathered on the subject of the assessment. No matter what the focus of the study, there is an infinity of indicators that can provide information about the action or subject under consideration.

A meaningful evaluation depends to a large degree upon the quality of the indicators utilized. By knowing the capacity of the indicators to assess the subject and the correlation between various indicators, it is possible to form inferences from one indicator to another. If the indicators are highly correlated, then it is no longer a question of which is best but which is least costly in time and/or resources. On the other hand, a total lack of correlation between indicators indicates one of two things: they are not measuring the same things, or at least one of them is a poor measure of the subject.

Here is a simple example of the concept of multiple indicators. There is a multitude of ways to determine the body fat percentage of a given person. One way is to simply derive the ratio of height to weight and apply the result to a table. If the gender of the person is known, the indicator is considerably more precise. If the circumference of the wrist is then added to the equation, the estimate is further refined. Including a skin fold measurement brings the estimate even closer to the actual body fat percentage of the subject. If the subject is weighed while immersed in a pool of water, the best possible approximation achievable without endangering the subject will be attained.

The correlation between these measures can easily be determined, and if each indicator is tested against the actual percentage of body fat as determined by the most accurate means available, it is possible to estimate the error of measurement of each of the less precise means. Once each indicator is optimized within the constraints of its own format, a decision can be made as to which single measure or group of measures best addresses the problem at hand.

If the best indicator is 99.1% accurate, the second best is 97% accurate, the third best is 93% accurate, and the fourth is 86% accurate, and they are all highly correlated, the decision as to which indicator to use is a function of needed accuracy versus cost/feasibility.

Absolute accuracy in any type of measurement is impossible. Disregard of cost--time, resources, impact upon our subject--is irresponsible. Responsible assessment entails careful consideration of these factors in light of the purpose of the assessment at hand.

In education, the assessment of learning is generally built around the demonstration of competence in certain domains. These domains and the goals and objectives that address them are codified in curricular frameworks and course outlines. Teachers design a course ofinstruction based upon these guidelines and, generally, even though teaching may take any of a number of forms, there is a high correlation between what is taught and the formal curriculum. It is this correlation that makes large-scale assessment possible. Indicators can be developed that measure learning along the articulated curriculum, and, because of the correlation between instruction and the stated curriculum, inferences can be drawn about the effectiveness of instructional strategies in school systems, schools, and classrooms along curricular lines.

The means used to determine whether students have achieved the goals set for them--the indicators employed--range from the results of simple observation to group-administered

standardized tests; from a short homework assignment to the performance and analysis of a complex laboratory experiment. Test scores, performance, and artifacts are all indicators of learning. Determining which indicators are best suited to specific purposes is the core of the assessment debate.

The Properties of Assessment Methodologies

Coverage

Advocates of alternative assessment assume that standardized tests provide a limited and shallow description of what an individual knows. In fact, multiple-choice tests cover a broad range of questions in each field surveyed, thereby providing a more detailed picture of student learning than is implied by the previous statement. Linn, Baker, and Dunbar (1991) state that ". . . breadth of content . . . may be one of the criteria by which traditional tests appear to have an advantage over more elaborate performance assessments. . . . [I]t is one of the criteria that clearly needs to be applied to any assessment." (p. 20). These authors further contend that lack of adequate coverage can lead to a distortion of the instruction provided and abnormally high scores. Feinberg (1990) points out that "compared to multiple-choice tests of similar length, written exams more arbitrarily emphasize one topic or another with which a student may (or may not) be familiar" (p. 17).

Even though a student's understanding of the tasks assessed can be examined in great detail with some alternative assessment techniques, the scope of the assessment is limited by the very constraints that comprise its raison d'etre. Alternative assessment, because of time constraints and, in the case of performance assessments, the complexity of the exercises, is generally limited to only a few tasks (Maeroff, 1991, p. 277). This is a problem. In a study of science performance assessment, Schavelson, Baxter, and Pine (1992) found that "task-sampling variability is considerable. In order to estimate the student's achievement, a substantial number of tasks may be needed." (p. 26). Wiggins notes that "writing prompts and performance situations in general are quite particular. What happens when we slightly vary the prompt or the context? One of the unnerving findings is: the student's score changes" (Brandt, 1992, p. 36).

What this means is that there is limited ability to generalize from task to task. Put another way, "Shavelson et al. found that performance was highly task dependent. The limited generalizability from task to task is consistent with research in learning and cognition that emphasizes the situation and context-specific nature of thinking" (Linn, Baker, & Dunbar, 1991, p. 19). Achieving a "substantial number of tasks" in alternative assessment is often impossible, making a generalization from the performance tasks to the larger realm of the subject area being assessed problematic. As Wiggins points out, "unpiloted, one-event testing in the performance area is even more dangerous than one-shot multiple-choice testing, because multiple-choice tests have many different but related items, which makes reliability easier to get and measure" (Brandt, 1992, p. 36).

Time

In Britain, where an assessment system based on students' performance on "standard assessment tasks" (SATs), a series of individually administered, performance-based measures, has been piloted in the areas of English, math, and science, "it has been estimated that the SATs took two to five weeks out of the school year" (Madaus & Kellaghan, 1993, p. 467). Nuttall states that, although the SATs had several good effects, "the interruption of normal education was substantial, as were the extra costs incurred." As a result, he continues, "at Grade 2, . . . the tasks have been redesigned so that many can be administered to the whole class at the same time,

and some of the most time-consuming ones have been dropped" (1992, pp. 57-57). In fact, the number of tasks students are expected to perform will be only about a third of the number originally proposed, and there is a possibility that multiple-choice questions will be used to further speed up the assessment (Maeroff, p. 279). Finally, Madaus and Kellaghan report that according to a study carried out by Patricia Broadfoot and her colleagues, "virtually all the teachers surveyed. . . reported that major disruptions had occurred to normal classroom practice, and half of those surveyed felt that the SATs were totally unmanageable" (p. 463). At a time when teachers are demanding more time to teach, will they buy in to an assessment program that requires such a vast amount of time be diverted from instruction?

Cost

Hymes et al., in the American Association of School Administrators (AASA) Critical Issues Report, The Changing Face of Testing and Assessment; Problems and Solutions (1991), offer the following insight into the problem of cost of assessment models:

> In these days of shrinking budgets, the cost-effectiveness of nationally standardized tests is a major boon to most local school districts. They can, in effect, get accountability for pennies a pupil. The alternatives are far more expensive. Even while defending their programs as well worth the investment, such pioneers in authentic assessment measures as Dale Carlson, director of the California Assessment Program, will concede that the cost differential can be as much as five times per pupil." (p. 11).

As Lorrie Shepard notes in Educational Researcher, "cost is a big factor, both for development and scoring" of alternative assessment tasks (1991, p. 22). Worthen (1993) credits Shepard with estimating the cost of the fourth grade math portion of the National Assessment of Educational Progress (NAEP) at $150 per pupil. Further, he tells us that

> George Madaus has reported estimates that using performance assessment in most subjects in American schools would cost between $2.5 and $3 billion annually. And Desmond Nuttall has pointed out that, despite several advantages of the broad use of performance assessment in England, its financial and personnel costs are so immense as to threaten its continuation (p. 452).

Other experts are no more hopeful in their estimates:

> For his estimate, Arthur Wise used the College Entrance Examination Board's fee for Advanced Placement (AP) tests: $65 per test. His projection of costs for five tests at three grade levels comes to between $2 billion and $3 billion a year. Keep in mind that the AP tests for the most part are scored by machine. . . . In several [European] countries it is estimated that to score essay-on-demand exam papers in four to five subject fields at ages 16+ and 18+ costs $135 per student (Madaus & Kellaghan, 1993, p. 467)

Testing the Few as Opposed to Testing the Many

To mitigate the cost of alternative assessment, Shepard (1991, p. 22) suggests the use of a sampling of students and grades in key subjects through the use of "a few exemplary assessments" as opposed to the universal testing of most students in several subjects and in several grades through many assessment items, as is normally the case with standardized testing.

This trade-off places severe limitations on the uses to which the assessment results can be put.

If Shepard's suggestion was implemented, the resulting assessment data would be practically useless for assessment of educational entities. As Madaus and Kellaghan tell us, "matrix sampling offers one way to reduce costs -- but only at the expense of information on individual performance. It is difficult to imagine that parents will be satisfied with the assurance that 'everything is progressing nicely although we don't know exactly how your child is doing'" (1993, p. 467). Furthermore, whereas deducing the progress of a cohort of students on the basis of limited data can be logically accomplished through time-honored statistical means, it cannot be done unless the information is generalizable. Discerning the effectiveness of any individual teacher, school, or program would be nearly impossible due to the inadequacy of the non-standardized data to provide generalizable information.

For these reasons, unless the cost of alternative assessment can be mitigated to allow for much more wide-spread implementation, its use as a model for large-scale educational assessment is problematic, at best.

Effects of High-Stakes Consequences on Assessment Measures

The British Educational Research Association has concluded that the validity of the Standard Assessment Tasks, a performance-based assessment system by which student achievement is assessed in Britain, may be compromised by teachers coaching their students because of the high stakes consequences of the results (Madaus & Kellaghan, p. 467). Standardized tests may be compromised in the same manner. In either case, the instructional process itself is subverted when assessment results become the goal of instruction.

Fresh, non-redundant, equivalent tests, regardless of the format, are the simplest means to discourage "teaching to the test." Revising standardized tests by drawing from item banks that have been constructed based on item response theory is a simple and inexpensive matter. In Tennessee and in a growing number of other states, such revisions are now mandated for tests administered state-wide. For alternative assessment schemes, which require that criteria for acceptable performance, articulation of performance levels, and training of assessors be revised in addition to devising a new problem or performance task, revision is clearly more difficult, time-consuming, and expensive. Furthermore, the complex task assessments are themselves difficult to devise. For these reasons, alternative assessment tasks are often administered unchanged, year after year.

In New York, where it is the case that the same tasks have been used to assess fourth-grade science manipulative skills for at least three years, the teachers are fully aware of what their students will be asked to do, so their students' performance on these tasks may actually reflect the results of practice rather than any higher-level cognition, thereby subverting the very purpose of alternative assessment (Maeroff, 1991, p. 281). Linn, et al., report the case of a New York geometry teacher

> who had been recognized for superior teaching on the basis of performance of his students on the Regents geometry exam. Unfortunately, the superior performance of his students was achieved by having them memorize the 12 proofs that might appear on the Regents exam. (p. 20)

It is apparent from cases such as these that we cannot assume that basing high-stakes assessment on alternative assessment would alleviate "teaching to the test." Rather, unless continual revision were an integral aspect of the process, the complexity of revising assessment tasks could exacerbate the problem.

Bias

Norm-referenced testing has confronted the problem of bias for decades. In fact, Feinberg notes that "multiple-choice tests themselves came into widespread use over the past four decades partly in an effort to achieve the very fairness that the critics say they lack" (p. 14), and on the next page, he cites Stephen P. Klein of the RAND Corporation who says that "if the performance test scores were adjusted to take into account the lower reliability of performance test grading, the racial gap would be even wider than on multiple-choice." Maeroff notes that

> In England in the late 1980s, when the assessments that make up the General Certificate of Secondary Education were changed to put more emphasis on performance tasks (which are assessed by classroom teachers) and less on written answers, the gaps between the average scores of various ethnic groups increased rather than narrowed. (p. 281).

Kellaghan et al. observed that teachers in Irish primary schools were quite biased in the evaluation of their students at the time of their study. They speculate that the reason such bias existed was due to the lack of standardized testing in Ireland (Kellaghan, Madaus, & Airasian, 1982, p. 23). Worthen and Spandel (1991, p. 67) put it this way:

> Assume for the moment that there is a bit of cultural bias in college entrance tests. Do away with them, right? Not unless you want to see college admission decisions revert to the still more biased "Good Old Boy" who-knows-whom type of system that excluded minorities effectively for decades before admissions tests, though admittedly imperfect, provided a less biased alternative."

Bias in standardized testing can be detected and, when it cannot be eliminated, its effects can be measured so that scores can be fairly interpreted. Bias in alternative assessment is much more difficult to articulate. Because this is the case, the effects of biased non-standardized assessments may not be recognized as such and may, therefore, be attributed to the subject rather than to the assessment.

Assessment and Cognitive Complexity

Perhaps the most common argument proponents of alternative assessment bring against the use of standardized tests is this: standardized tests measure only recall and other lower-order thinking skills whereas alternative methods of assessment require students to exhibit the higher-order skills such as critical thinking, analysis, synthesis, reasoning, and problem solving. If this were true, it would be a very damning argument indeed, but neither assertion is altogether accurate.

The development of assessments that require the demonstration of complex problem-solving strategies is an essential component of the alternative assessment movement. Proponents of alternative assessment cite the ability of alternative assessment problems to elicit the use of higher-order thinking skills and to assess the quality of those skills as its primary advantage over standardized methods of assessment.

While it is true that alternative assessment techniques certainly have the capacity to fulfill these high aspirations, it does not necessarily follow that they always do. Worthen (1993, p. 450) points out that "proponents of alternative assessment cannot assume that students are using such skills just because they are performing a hands-on task." Linn et al. address this problem in more detail:

> The construction of an open-ended proof of a theorem in geometry can be a cognitively complex task or simply the display of a memorized sequence of responses to a particular problem, depending on the novelty of the task and the prior experience of the learner. (p. 19)

Performance assessments may show how students solve problems or how much they have practiced the skill being assessed. Essays may indicate the ability to organize thoughts and communicate them in writing or nothing more than the acquisition of a formula for writing essays, assiduously taught in preparation for the assessment. Discerning the difference is another hurdle alternative assessment must surmount.

As to the assertion that standardized, multiple-choice tests assess only recall of specific and isolated bits of knowledge, it is appropriate to admit that some do. Some criterion-referenced tests of basic skills are little more than recitations of factoids. But this is not a function of standardized tests: this is a function of the purpose for which a particular test was devised. As Worthen and Spandel put it,

> The notion that multiple choice tests can tap only recall is a myth. In fact, the best multiple choice items can--and do--measure students' ability to analyze, synthesize information, make comparisons, draw inferences, and evaluate ideas, products, or performances. (p. 67)

Feinberg points out that "the most widely known multiple-choice exam, the Scholastic Aptitude Test, tests very little knowledge; it is almost completely a test of analytical and reasoning ability at quite complex and sophisticated levels" and that "many other standardized tests, particularly at the high school level, also probe the ability to draw fair inferences and reach tenable conclusions." (p. 16). Although open-ended alternative assessments offer students a far greater range for response than any multiple-choice format can, it is not accurate to assert that higher-order thinking skills cannot be assessed with standardized tests.

Finally, in an article he wrote for *Educational Leadership*, Whimbey (1985) found that there was a high correlation between aptitude and achievement test scores and the scores on special reasoning tests. He concludes that "the high correlations mean that there is generally little value in administering a separate reasoning test like the WASI (Whimbey Analytical Skills Inventory) if scores on a battery of aptitude and achievement tests such as the NJCBSPT (New Jersey College Basic Skills Placement Test) are already available for students." (p. 38).

Standardization of Assessment Situations

Standardization is of little importance if the results of assessment are to be used in isolation from all other factors. In other words, if the purpose is simply to learn about the state of a single subject, a unique assessment might be devised to furnish the information desired. However, if the assessment is to be used for the purpose of comparison, generalization, or decision-making, standardization is essential.

Standardized testing achieves standardization by norming practices, machine scoring of multiple-choice questions, precise instructions for administration, and standard formats for tests and recording of responses. The results can then be used to draw inferences about the state of cohorts or individuals as compared to an established standard.

The task of standardization is far more complex in the matter of alternative assessment. The judgment of performance of designated tasks is a matter of interpretation and is carried out by any number of individuals who may have different understandings of what an appropriate response entails. The presentation of the task may take place in vastly different circumstances

and contexts. Madaus and Kellaghan reported major problems with standardization of assessment procedures and settings in the British system of performance assessment, the SATs:

> . . . there was a serious lack of standardization, which must call into question the comparability of individual student scores or aggregate school scores. . . . [T]he report from the British Educational Research Association (BERA) concluded that assumptions that the SATs produce reliable and robust data were not borne out.
>
> The lack of comparable, reliable, and robust data is a function of a lack of standardization in the administration of the SATs, problems in making judgments about students' performance, and of wide variations between schools in the support received for testing and in the amount of changes in practices and routines occasioned by the assessments. (p. 466).

Furthermore, those who are assessed may also have difficulty in understanding what response is needed. Maeroff cites problems in portfolio assessment of elementary students' work. In one case, when a student chose her "best" math work, it was all from the beginning of the school year. The evaluator had nothing from which to judge the child's progress (p. 280).

It is not necessary that alternative assessment measures exhibit the same level of standardization as achievement tests do in order to be of value, but the appropriate use of alternative assessment may be limited by the very characteristics that make it a good indicator of individual achievement from being a valid indicator of how that progress compares to the progress of others.

> It may be that an alternative assessment that is a marvelous indicator of an individual child's academic progress will prove fairly useless for other purposes. Americans may have to decide whether comparisons are what they seek in alternative assessment or whether they prefer to use the approach for other, more individualized purposes. . . . Putting less emphasis on comparisons is fine, but at some point a child and his parents have a right to know whether the child's progress is reasonable for his or her age and experience. (Maeroff, p. 276).

Is There Evidence of Correlation Between Alternative and Standardized Assessment Measures?

Shavelson et al. found that "taken in the aggregate, a combination of the alternative assessments correlates about the same with aptitude as does the standardized science achievement test. Aptitude, then, is a major factor in generalizing performance across assessment tasks" (p. 26). What this means is that students score at a comparable level on the battery of performance assessments in Shavelson's study and on aptitude and a standardized science achievement test.

Feinberg cites further studies in which a correlation was found between performance testing and standardized measures:

> On the California Bar exam, the largest program so far to have incorporated performance testing, the rank order of applicants is nearly the same on the performance, essay, and multiple-choice sections. Low-scoring students score low on all three parts, high-scoring candidates score high on all three.
>
> According to a new study, the same thing is true on the free-response and multiple-choice parts of the Advanced Placement computer science exam. Several similar studies on other tests have yielded similar conclusions. . . . (p. 31)

Whimbey's work in this area has been cited previously in the section, "Assessment and Cognitive Complexity."

More work examining the correlation between alternative assessment ratings and standardized measures of learning must be done, but there is evidence that such a correlation exists. Since this is the case, the question becomes which form of assessment is most appropriate to a given purpose. Having determined that, we must next ask which appropriate model is most efficient in terms of cost and resources.

The Tennessee Value-Added Assessment System (TVAAS): Recent Refinements in the Use of Standardized Student Achievement Data in Educational Assessment

What TVAAS Is.

TVAAS is a process that measures the influence that systems, schools, and teachers have on the rate of academic growth for populations of students. To accomplish this, TVAAS uses statistical mixed-model methodology and student scale scores from the norm-referenced component of the Tennessee Comprehensive Assessment Program (TCAP). However, any reliable linear measure of academic growth with a strong relationship to the curriculum could be used as input into the process. Pilot studies revealed and subsequent research has confirmed that the statistical mixed model theory and methodology upon which TVAAS is based alleviates the problems associated in the past with the use of student data in educational assessment (McLean & Sanders, 1984). When extraneous factors are identified that would bias the estimates, this methodology readily allows the incorporation of exogenous variables to insure unbiased results. However, the research to date indicates that this has not been necessary. For example, the effects attributable to individual systems and schools have been shown to be unrelated to socio-economic indicators such as number of reduced-cost or free lunch students, racial composition of the student body, or location--rural, urban, suburban. Very simply put, each child's own academic history incorporates socio-economic status, ability, past achievement, and many other factors. By modeling a learning profile for each student as part of the mixed-model equations, children serve as their own controls or "blocking factors" in TVAAS.

Assessment should be a tool for educational improvement, providing information that allows educators to determine which practices result in desired outcomes and which do not. TVAAS is an outcomes-based assessment system. By focusing on outcomes rather than the processes by which they are achieved, teachers and schools are free to use whatever methods prove practical in achieving student academic progress. TVAAS does not assume a "perfect teacher" or a "best way to teach." Rather, the assumption is that effective teaching, whatever form it assumes, will lead to student gains. The advantage TVAAS offers in this regard is that those teachers and methods that lead to greater student achievement can be identified. In other words, teachers can try something new and actually see the effects, at least insofar as they are reflected in student academic gains.

In several ways, this is an entirely new approach to using normed data. One criticism of the use of normed data is that it is often used to place students somewhere on a distribution for the purpose of comparison with others. TVAAS, to the contrary, uses scale scores to establish where a child is academically and to determine how much progress that child makes in a subject year. Where s/he is in relation to other students is irrelevant. Whether s/he progresses normally from whatever point s/he begins is what matters. TVAAS concentrates on gains because student gains provide information on educational effects that measures of ability cannot. High achievement scores do not necessarily indicate progress, but high gains do. By focusing on the gains that all students make from year to year, regardless of where they start, the school systems and the individual schools deemed to be most effective by TVAAS are those that provide educational

opportunities for all students--the advanced learner as well as the slower learner.

Astin (1982, p. 14) states that "the basic argument underlying the value-added approach is that true excellence resides in the ability of the school or college to affect its students favorably, to enhance their intellectual development, and to make a positive difference in their lives." TVAAS was developed on the premise that society has a right to expect that schools will provide students with the opportunity for academic growth regardless of the level at which the students enter the educational venue. In other words, all students can and should learn commensurate with their abilities.

This very brief description of TVAAS is provided simply to illustrate one viable model for utilizing standardized scores in assessing educational entities. For a far more detailed description of TVAAS, its development, methodology, and application, see Sanders & Horn, 1994.

Conclusion

Standardized testing renders viable, inexpensive, reliable, and valid indicators of student learning useful in the assessment of educational entities and student achievement. Standardized testing is already in place in most states of the union, so the data are readily available. Standardization makes it possible to generalize and to draw conclusions about the data and their implications.

Alternative forms of assessment are also viable tools for the assessment of student progress and attainment, so long as care is taken to assure their validity and reliability. Because they are expensive and difficult to develop, administer and score, their usefulness for large-scale assessment is questionable. Should such assessment models achieve comparable reliability and validity as standard measures now possess, they would in effect have become standardized, also. There is a question as to whether this is a desirable result. If it is determined that this is the course that should be attempted, the results of assessments of any type that exhibit an appropriate degree of reliability and validity can be used for large-scale assessment. However, if alternative forms of assessment are used, instead, for the assessment of individual students by in-house assessors, many of the problems listed above may be avoided and the strengths of alternative assessment modes may have the impact desired on the quality of instruction in the classroom.

The issue is not whether one form of assessment is intrinsically better than another. No assessment model is suited for every purpose. The real issue is choosing appropriately among indicator variables and applying the most suitable model to render them. It is necessary to determine what information is sufficient to each purpose before deciding upon the form of assessment to be used. When a variety of valid and reliable assessment methods exist, it is parochial and ineffectual to adhere to only one, asserting that it is in all instances superior. It is the opinion of these authors that factionalism is detrimental to the comprehension of educational effects and that much is to be gained by adopting a more ecumenical stance in regard to educational assessment.

References

Astin, A. W. (1982). "Excellence and Equity in American Education." Washington, D. C.: National Commission on Excellence in Education. (ERIC Document Reproduction Service No. ED 227 098)

Brandt, R. (1992). "On Performance Assessment: A Conversation with Grant Wiggins." *Educational Leadership*, 49(8), pp. 35-37.

Feinberg, L. (1990). "Multiple-Choice and Its Critics." *College Board Review*, No. 156, pp.

12-17+.

Hymes, D., A. Chafin, & P. Gonder. (1991). The Changing Face of Testing and Assessment; Problems and Solutions (AASA Critical Issues Report). Arlington, VA: American Association of School Administrators.

Kellaghan, T., G. Madaus, & P. Airasian. (1982). *The Effects of Standardized Testing*. Boston: Kluwer-Nijhoff Publishers.

Linn, R. L., E. L. Baker, & S. B. Dunbar. (1991). "Complex, Performance-Based Assessment: Expectations and Validation Criteria." *Educational Researcher*, 20(8), pp. 15-21.

Maeroff, G. I. (1991). "Assessing Alternative Assessment." *Phi Delta Kappan*, 73(4), pp. 273-281).

MacLean, R. A., and Sanders, W. L. (1984). Objective Component of Teacher Evaluation--A Feasibility Study, (Working Paper No. 199). Knoxville, TN: University of Tennessee, College of Business Administration.

Nuttall, D. (1992). "Performance Assessment: The Message from England." *Educational Leadership*, 49(8), pp. 54-57.

Sanders, W. L. & S. P. Horn. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment." *Journal of Personnel Evaluation in Education*, 8, pp.299-311.

Shavelson, R., G. Baxter, & J. Pine. (1992). "Performance Assessments: Political Rhetoric and Measurement Reality." *Educational Researcher*, 21(4), pp. 22-27.

Shepard, L. (1991). "Interview on Assessment Issues With Lorrie Shepard." *Educational Researcher*, 20(2), pp. 21-23+.

Whimbey, A. (1985). "You Don't Need a Special 'Reasoning' Test to Implement and Evaluate Reasoning Training." *Educational Leadership*, 43(2), pp. 37-39.

Worthen, B. (1993). "Critical Issues That Will Determine the Future of Alternative Assessment." *Phi Delta Kappan*, 74(6), pp. 444-454.

Worthen, B. & V. Spandel. (1991). "Putting the Standardized Test Debate in Perspective." *Educational Leadership*, 48(5), pp. 65-69.

About the Authors

William L. Sanders
Sandra P. Horn
sphorn@sacam.oren.ortn.edu
Value Added Research & Assessment Center
University of Tennessee

EPAA can be accessed either by visiting one of its several archived forms or by subscribing to the LISTSERV known as EPAA at LISTSERV@asu.edu. (To subscribe, send an email letter to LISTSERV@asu.edu whose sole contents are SUB EPAA your-name.) As articles are published by the *Archives*, they are sent immediately to the EPAA subscribers and simultaneously archived in three forms. Articles are archived on *EPAA* as individual files under the name of the author and the Volume and article number. For example, the article by Stephen Kemmis in Volume 1, Number 1 of the *Archives* can be retrieved by sending an e-mail letter to LISTSERV@asu.edu and making the single line in the letter read GET KEMMIS V1N1 F=MAIL. For a table of contents of the entire ARCHIVES, send the following e-mail message to LISTSERV@asu.edu: INDEX EPAA F=MAIL, that is, send an e-mail letter and make its single line read INDEX EPAA F=MAIL.

The World Wide Web address for the *Education Policy Analysis Archives* is http://seamonkey.ed.asu.edu/epaa

Education Policy Analysis Archives are "gophered" at olam.ed.asu.edu

To receive a publication guide for submitting articles, see the *EPAA* World Wide Web site or send an e-mail letter to LISTSERV@asu.edu and include the single line GET EPAA PUBGUIDE F=MAIL. It will be sent to you by return e-mail. General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, Glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692)

Editorial Board

Dennis Sayers
dmsayers@ucdavis.edu

Jay Scribner
jayscrib@tenet.edu

Robert Stonehill
rstonehi@inet.ed.gov

Robert T. Stout
stout@asu.edu