NISE BRIEF

REPORTING ON ISSUES AND RESEARCH IN SCIENCE, MATHEMATICS, ENGINEERING, AND TECHNOLOGY EDUCATION

Value-Added Indicators: A Powerful Tool for Evaluating Science and Mathematics Programs and Policies

by Robert H. Meyer

States and districts are increasingly turning to school accountability as an instrument of reform.

Educational outcome indicators frequently are used to measure the performance of schools, programs, and policies. Reliance on such indicators is largely the result of a growing demand to hold these entities accountable for their performance, defined in terms of outcomes, such as standardized test scores in mathematics, science and reading, rather than inputs, such as teacher qualifications, class size, or the quality of lab facilities. This Brief discusses the weaknesses of the most commonly used educational outcome indicators—average and median test scores and proficiency-level indicators—and the advantages of value-added indicators.[1] Several major conclusions emerge from the analysis.

First, the most common educational indicators are highly flawed as measures of school and program performance, even if they are derived from highly valid assessments. As a result, they are of limited value, if not useless,

The most common educational indicators—average and median test scores and proficiency-level indicators—are highly flawed as measures of school and program performance.

for evaluating relative program performance or program performance over time and thus should not be used to hold schools or programs accountable for their performance. Simulation results indicate that changes over time in average test scores could very well be negatively correlated with actual changes in program performance.

Second, the typical indicators used to assess school and program performance provide institutions with the perverse incentive to "cream," that is, to raise measured performance by educating only those students who tend to have high test scores. The potential for creaming is apt to be particularly strong in environments characterized by selective admissions. However, creaming could also exist in subtler, but no less harmful, forms. For example, schools and programs could create an environment that is relatively unsupportive for potential dropouts, academically disadvantaged students, and special education students, thereby encouraging these students to drop out, transfer to another school, or enroll in a different program. Second, schools could aggressively retain students at given grade levels. Finally, high quality teachers and administrators could gravitate to schools and programs that predominantly serve high-scoring students.

Third, typical performance indicators tend to be biased against schools and programs that disproportionately serve academically disadvantaged students. One source of bias is the well-known fact that school productivity is only one of the many determinants of student achievement. Most of the variation in average or median test scores can usually be accounted for by differences in the types of students enrolled across schools and programs.

Finally, given the problem of student mobility (as well as several other problems discussed in the Brief), it is not possible to construct statistically valid performance indicators if tests, assessments, or other student outcomes are measured so infrequently that a significant proportion of students change schools between testings.

Given the substantial problems that exist with common educational outcome indicators, what should be done to improve the situation? If one is interested in having indicators that are appropriate for accountability and/or evaluation purposes, a solution is to design indicator/evaluation systems based on the value-added approach, as has recently been done in a number of states and districts, for example, Dallas, Minneapolis, South Carolina, and Tennessee.

The essence of the value-added approach is that school and program performance is measured using a statistical regression model that includes, to the extent possible, all of the nonschool factors that contribute to growth in student achievement, in particular, prior student achievement and student, family, and neighborhood characteristics. The key idea is to statistically isolate the contribution of schools and programs to *growth in student achievement* at a given grade level from all other sources of student achievement growth.[2] This isolation is particularly important in light of the fact that differences in prior achievement and student and family characteristics account for far more of the variation in student achievement than school-related factors. Failure to account for differences across schools in these characteristics could result in highly contaminated indicators of performance.

A Critique of the Average Test Score as a Measure of School and Program Performance

The critique applies to any indicator that measures some aspect of the level of student achievement, including, for example, median test scores and proficiency-level indicators. A school-level average test score is a highly flawed measure of school performance for four basic reasons. One, the average test score is *contaminated by factors other than school performance*, in particular, the average level of student achievement prior to entering first grade — average initial achievement—and the average effects of student, family, and community characteristics on student achievement growth from first grade through the grade in which students are tested. In fact, it is quite likely that comparisons across schools of average test scores primarily reflect these differences rather than genuine differences in intrinsic school performance. Thus, average test scores are highly biased against schools that disproportionately serve academically disadvantaged students and communities.

Two, the average test score reflects information about school performance that tends to be grossly *out of date*. Consider, for example, the average test score

Measuring a school's contribution to student achievement requires every year testing.

for a group of students tested at the end of tenth grade. The average test score for this group is a reflection of the accumulated learning that occurred in tenth grade during the prior year, two years earlier in ninth grade, three years earlier in eighth grade, and so on, all the way to kindergarten and preschool—eleven (or more) years earlier. Indeed, a tenth-grade level indicator could be dominated by information that is five or more years old. One consequence is that changes over time in average test scores could be negatively correlated with actual changes in program performance (Meyer, 1996). The fact that average test scores reflect out-of-date and possibly misleading information severely weakens them as instruments of public accountability. To allow educators to react in a timely and responsible fashion, performance indicators must reflect information that is current and accurate.

Three, average test scores at the school, district, and state levels tend to be highly *contaminated due to student mobility*. For example, the typical high school student is likely to attend several different schools over the period spanning kindergarten through twelfth grade. For these students, a test score reflects the contributions of more than one and possibly many different schools. The problem of contamination is compounded by the fact that rates of student mobility tend to differ dramatically across schools. Contamination is apt to be especially high in communities that undergo rapid population growth or decline and in communities that experience significant changes in their occupational and industrial structure. Contamination due to student mobility is probably a relatively minor problem at the national level, since rates of in- and out-migration are low compared to rates of mobility within the nation, but at the district and school levels it is apt to be quite serious.

Finally, the average test score fails to *localize* school performance to a specific classroom or grade level—the natural unit of accountability in a traditional school. This lack of localization is, of course, most severe at the highest grade levels. A performance indicator that fails to localize school performance to a specific grade level or classroom is likely to be a relatively weak instrument of public accountability.

An Example Based on National Data

The practical significance of the above analysis is illustrated using data on average mathematics scores from 1973 to 1986 from the National Assessment of Educational Progress (NAEP).[3] As indicated in Panel A of Table 1, NAEP scores for eleventh grade exhibit the by-now familiar pattern of sharp declines from

1973 to 1982 and then partial recovery between 1982 to 1986. The eleventh-grade data, by themselves, are fully consistent with the premise that academic reforms in the early and mid 1980s generated substantial gains in academic achievement. In fact, an analysis of the data based on a gain indicator (a value-added type indicator) rather than an average test score suggests the opposite conclusion—see Panel B of Table 1.

The gain indicator is similar to a true value-added indicator in that it controls for differences among students in prior achievement. It does so in a very simple and intuitive way: gain is the change in average test scores over time (and across grades) for the *same cohort* of students. For example, the gain in test scores for students who were eleventh-grade students in 1986 is given by average test score of eleventh-grade students in 1986 minus the average test score for seventh-grade students in 1982 (four grades and four years earlier) (that is, 302.0 – 268.6 = 33.4). Unfortunately, the gain indicator, unlike the value-added indicator, does not control for differences in student, family, and neighborhood characteristics that contribute to growth in student achievement. As a result, the gain indicator reflects possible changes over time in the composition of the population as well as changes in school productivity.[4] Nonetheless, it is instructive to compare the gains in achievement experienced by different cohorts.[5]

As indicated in Panel B, the achievement growth of high school students (from seventh to eleventh grade) during the 1982 and 1986 period was actually no better than achievement growth during previous periods. In fact, the gain from seventh to eleventh grade was actually slightly lower during the 1982 to 1986 period than in previous periods! The rise in eleventh-grade math scores from 1982 to 1986 stems from an earlier increase in achievement growth for that cohort rather than from an increase in achievement growth over grades seven to eleven. In short, these data provide no support for the notion that high school academic reforms generated significant increases in test scores during the mid-1980s. These data also vividly confirm the general superiority of the gain indicator, relative to level indicators such as the average test score, as a measure of educational productivity.

It would be interesting to report the above analysis using true value-added as opposed to gain indicators. Unfortunately, the NAEP data do not permit such an analysis to be conducted, since the same students are not sampled for two consecutive NAEP surveys. This weakness in NAEP data could be remedied by switching to a survey design that was at least partially longitudinal.

Value-Added Indicators: Data Requirements

Given the problems that exist with the average test score and other level indicators and, to a lesser degree, the gain indicator, it is important to consider whether value-added indicators could potentially be used as the primary tool for evaluating the performance of schools and programs. There are at least two reasons to be optimistic in this regard. First, value-added models have been used extensively over the last three decades by evaluators and other researchers interested in education and training programs. Second, a number of districts and states, including Dallas, Minneapolis, South Carolina, and Tennessee, have successfully implemented value-added indicator systems.[6]

Nonetheless, despite the promise of value-added indicator systems, it is clear that they require a major commitment. In particular, districts and states must be prepared to (1) assess students frequently and (2) develop comprehensive district or state data systems that contain information on student test scores and student, family, and community characteristics. The need for frequent testing stems from the fact that value-added indicators are designed to measure the contribution of schools to growth in student achievement over a given time period. In order to be able to construct

Table 1. NAEP Mathematics Examination Data

(A) Average Test Scores by Year

GRADE	1978	1978	1982	1986
3rd	219.1	218.6	219.0	221.7
7th	266.0	264.1	268.6	269.0
11th	304.4	300.4	298.5	302.0

(B) Average Test Score Gain From Year to Year for Each Cohort

GRADE	1973 to 1978	1978 to 1982	1982 to 1986
3rd to 7th	45.0	50.0	50.0
7th to 11th	34.4	34.4	33.4

Source: Dossey et al. (1988).

value-added (or gain) indicators it is therefore necessary to have achievement data for the same individuals at two points in time. Students who are missing either pre- or posttest data must be excluded from the analysis and thus from a district's accountability and/or evaluation system.

From the perspective of measuring school and program performance, an ideal testing program would do the following:

- Test all students annually during the late spring. Many districts currently follow this practice.
- Test all students who attend summer school at the end of the summer (or in the fall at the beginning of the subsequent school year). Following the recent boom in summer school enrollments, many districts have begun testing students at the end of summer school.
- Test mobile students at the point of entry into the district (or into a new school in the district).[7] Minneapolis is one of the districts that is pioneering the use of entry-point testing. As indicated below, this component is very important in a comprehensive assessment program.

Student and family characteristics also contribute to student achievement.

Annual testing has three major advantages. First, it maximizes accountability by localizing school and program performance to the most natural unit of accountability: the grade level or classroom. Second, it yields up-to-date information on performance. Third, it severely limits the number of students who would be excluded due to student mobility and, as a result, yields a data set that is likely to be highly representative of the school population as a whole and large enough to yield statistically reliable school performance estimates. On the other hand, less frequent testing, say testing at grades kindergarten, 4, 8, and 12, might be acceptable for national purposes, since student mobility is not really an issue at the national level. For purposes of evaluating local school and program performance, however, the problems created by student mobility argue strongly for frequent testing.

Adding a post-summer-school test yields one additional advantage; namely, it allows districts to separately evaluate the productivity of programs during the regular school year and those during the summer.[8] Adding a point-of-entry test for in-migrant students enables districts to evaluate the degree to which mobile students experience growth in achievement that is comparable to that of nonmobile students. Furthermore, it allows these students to be included in state and district performance indicators.[9] When schools are increasingly under pressure to achieve high (measured) performance, adopting an indicator/evaluation system that systematically excludes any group in the population seems particularly unwise.

One potential obstacle to producing high-quality value-added indicators is the difficulty of collecting extensive information on student and family characteristics. These data are required as "control variables" in value-added models. In most schools the following data are typically available from administrative records: race and ethnicity, gender, special education status, limited English proficiency (LEP) status, eligibility for free or reduced-price lunch, and whether a family receives welfare benefits. Supplemental surveys of students and parents may be used to collect other information, such as parental education and income and family attitudes toward education (variables known to be powerful determinants of student achievement growth).

The consequence of failing to control adequately for student, family, and community characteristics is that value-added indicators may be contaminated if there are major differences across schools and programs in unmeasured (uncontrolled) student, family, and community characteristics. Thus, value-added indicators derived from models with "weak" predictors of student achievement growth might be only slightly better than gain

indicators (better in the sense of being more highly correlated with a theoretically perfect value-added indicator). Even so, they are likely to be much better indicators than average test scores. The key issue, of course, is not whether a particular value-added indicator is perfect. Rather, the issue is whether the indicator provides a substantially better measure of school and program performance than other affordable indicators.

The cost of implementing an assessment system that is sufficient to support value-added (or gain) indicators is obviously higher than an assessment system that tests students only in selected grades (say, 4, 8, and 12). The thrust of this Brief is that an assessment system with infrequent testing is unlikely to produce outcome indicators that are valid for the purpose of measuring school performance. Thus, a district that is unwilling or unable to support the expense of frequent assessment should be very wary of using the achievement data that it does collect to evaluate the performance of schools and programs.

A value-added approach to school accountability is useful and possible.

Conclusions and Recommendations

Average and median test scores and proficiency-level indicators, the most commonly used indicators in American education, are highly suspect as indicators of school and program performance. These indicators suffer from four major deficiencies: they fail to localize performance to the classroom or grade level; they aggregate information on performance that tends to be grossly out of date; they are contaminated by student mobility; and they fail to measure the distinct contribution of schools and programs to growth in student achievement as separate from the contribution due to student, family, and community factors. As a result, they are flawed measures for evaluation purposes and are weak, if not counterproductive, instruments of public accountability.

The gain indicator (the change in average test scores from grade to grade for the same cohort of students) and the value-added indicator (the gain indicator statistically adjusted for differences across schools and programs in the type of students served) avoid the first of these four problems. In addition, the value-added indicator potentially eliminates the bias that exists in the gain indicator due to differences across schools in student, family, and community characteristics, particularly if it is based on a model that includes an extensive set of control variables. In this case, it fully eliminates the incentive for schools to cream.

The value-added approach to measuring school and program performance relies on a statistical model to identify the distinct contributions made by schools and programs to growth in student achievement. The quality of a value-added indicator is determined by four factors: the frequency with which students are tested, the quality and appropriateness of the tests that underlie the indicators, the adequacy of the control variables included in the value-added models, and the appropriateness (validity) of the statistical model used to used to define the indicator. In terms of the first factor, states and districts need to seriously consider testing students at every grade level, beginning with kindergarten; to further improve their indicator systems, states and districts need to think about testing summer school students and in-migrant students at the point of entry into the school or district. With respect to the second and third issues, it is important that states and districts make it a major priority to collect extensive and reliable information on student and family characteristics and to develop state tests that are technically sound and fully attuned to their educational goals. Finally, ongoing research is needed to assess the sensitivity of estimates of school and program performance to alternative statistical models and alternative sets of control variables.

ENDNOTES

[1] Proficiency-level indicators measure the proportion of students who score above a specified proficiency-level "cut point."

[2] Note that value-added indicators focus on the growth in student achievement from one grade to the next for given cohorts of students rather than on the change (or trend) over time in average test scores for students at a given grade level. Value-added indicators are thus based on longitudinal as opposed to cross-sectional student data.

[3] See Barton and Coley (1998) for a similar analysis that focuses on gains in student achievement for students age 9 to 13 from 1978 to 1996.

[4] The gain indicator also cannot be constructed if the before (pre) and after (post) tests differ and have not been placed on the same measuring scale.

[5] NAEP was originally designed to permit this type of analysis. In mathematics, the tests have generally been given every four years at grade levels spaced four years apart. For this illustrative analysis, we assume that average test scores in 1973 are comparable to the unknown 1974 scores.

[6] Millman (1997) contains detailed descriptions and analyses of the Dallas and Tennessee value-added systems.

[7] In principle, mobile students could also be tested prior to migrating out of a school or district. On the other hand, these students might not have much of an incentive to take a test just prior to leaving a school, and if they did take such a test, the results could be quite misleading. I do not see an easy way of including out-migrants in an accountability system other than testing all students at multiple points during the school year—an extremely expensive proposition.

[8] Optionally, all students—including non-summer-school students—could be tested in the late spring and early fall. The advantage of this approach is that it would allow schools to distinguish growth in student achievement during the school year from growth (or possibly decline) during the summer for all students. It would also allow schools to better estimate the benefits of participation in summer school. This approach would, of course, raise the costs of testing.

[9] In the absence of point-of-entry testing, mobile (in-migrant) students must be excluded from value-added or gain indicators because the students lack a prior measure of achievement.

REFERENCES

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade.* Princeton, NJ: Policy Information Center, Educational Testing Service.

Dossey, J. A., Mullis, I. V., Lindquist, M. M., & Chambers, D. L. (1988). *The mathematics report card: Are we measuring up?* Princeton, NJ: Educational Testing Service.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197-223). Washington, DC: National Academy Press.

FOR FURTHER READING

Bryk, A. S., & Raudenbush, S. W. (1989). Quantitative models for estimating teacher and school effectiveness. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 205-232). San Diego: Academic Press.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable* (pp. 23-63). Washington, DC: Brookings.

Hanushek, E. A., Taylor, L. (1990.) Alternative assessments of the performance of schools, *Journal of Human Resources*, *25*(2), 179–201.

Mandeville, G. K. (1994). The South Carolina experience with incentives. In T. A. Downes & W. A. Testa (Eds.), *Midwest approaches to school reform* (pp. 69-97). Proceedings of a conference held at the Federal Reserve Bank of Chicago, October 26-27.

Meyer, R. H. (1999). The effects of math and math-related courses in high school. In S. E. Mayer, & P. E. Peterson (Eds.), *Earning and learning: How schools matter* (pp. 169-204). Washington, DC: Brookings.

Millman, J. (1997). *Grading teachers, grading schools.* Thousand Oaks, CA: Corwin.

Raudenbush, S. W., & Willms, D. J. (1991). *Schools, classrooms, and pupils*. San Diego: Academic Press.

Raudenbush, S. W., & Willms, D. J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307-336.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299-311.

Willms, D. J., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, *26*, 209–232.

Robert H. Meyer is a Senior Scientist at the Wisconsin Center for Education Research at the University of Wisconsin-Madison and a Lecturer and Research Associate at the Harris Graduate School of Public Policy Studies at the University of Chicago.

The author would like to thank Andrew Porter, NISE Director; Adam Gamoran, University of Wisconsin-Madison; and Margaret Goertz, Consortium for Policy Research in Education, University of Pennsylvania, for very helpful comments and suggestions on this Brief. Many of the issues discussed in this Brief are considered at greater length in Meyer (1996).

Photos by Susan Lina Ruggles

This *Brief* was supported by a cooperative agreement between the National Science Foundation and the University of Wisconsin-Madison (Cooperative Agreement No. RED-9452971). At UW-Madison, the National Institute for Science Education is housed in the Wisconsin Center for Education Research and is a collaborative effort of the College of Agricultural and Life Sciences, the School of Education, the College of Engineering, and the College of Letters and Science. The collaborative effort also is joined by the National Center for Improving Science Education in Washington, DC. Any opinions, findings or conclusions herein are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

This publication is free on request. *NISE Briefs* are also available electronically at our World Wide Web site: **www.nise.org**

National Institute for Science Education
University of Wisconsin-Madison
1025 W. Johnson Street
Madison, WI 53706
(608) 263-9250
FAX: (608) 262-7428

E-mail: niseinfo@macc.wisc.edu

Vol. 3, No. 3 **June 2000**

Visit us at our World Wide Web site: **www.nise.org**

National Institute for Science Education
University of Wisconsin-Madison
1025 W. Johnson Street
Madison, WI 53706

Nonprofit Organization
U.S. Postage
PAID
Madison, Wisconsin
Permit No. 658