What Do Students Learn in High School?

Using Student Gain Scores to Evaluate Student Progress

Bruce Thompson

Rader School of Business

Professor of Management

Milwaukee School of Engineering

(414) 277-7378

FAX: (414) 277-7479

thompson@msoe.edu

Abstract

Using data from Milwaukee Public Schools, which tests students every year, this paper looks at both the trend in average scores from year to year and the average gains that students make from one year to the next. In ninth grade the average student makes no progress as measured by test scores in reading, language arts, and mathematics. This lack of progress is not unique to any single subgroup of students. I discuss possible causes for the ninth grade slump and suggest possible solutions.

What Do Students Learn in High School?

Using Student Gain Scores to Evaluate Student Progress

Introduction

Student achievement, especially among low-income minority students in urban schools, is a growing national concern. Reflecting this concern, passage of the federal No Child Left Behind Act will both increase the amount of testing in many schools and raise the associated stakes. Depending on both the value of their scores and the direction the scores take over time, schools will face a variety of consequences.

One limitation to the most common methods that use tests to rate schools is that they ignore both the social and economic challenges faced by the student body and the academic starting point of the students. Measuring a school's progress by comparing its average test scores in two years implies that the two groups of students are identical.

Proper use of data can largely overcome these limitations. Socioeconomic data can be incorporated through models that compare the performance of a school to that of other schools with similar populations. In Thompson (in press), I describe such a model.

The issue of varying student starting points can be approached by comparing a student's score in one year with that same student's score the previous year. Whatever their starting points, ideally all students would improve every year.

The terminology used to describe models that attempt to separate the value the school adds from student characteristics is often confusing and inconsistent. In some studies, measuring individual student progress is referred to as "longitudinal"assessment. Elsewhere, however, the same term refers to evaluations that compare schools' average scores from one year to the next using different student populations. In other studies, particularly those associated with Tennessee, comparing individual student scores over time is described as "value-added." However, other authors use the term to refer more generally to any model that tries to control for a school's inputs, including student socioeconomic status. In this paper, I use the term "gain score" when referring to the difference between the score of an individual student in one year and the same student's score in the previous year.

The use of gain scores generally requires that students be tested every year using comparable tests. This requirement probably helps explain why gain scores have not been adopted more widely. With the new federal law requiring annual testing in reading and mathematics, more states may adopt assessment plans that allow gain scores to be calculated.

Probably the most widely publicized use of gain scores has been the Tennessee "value-added" system (TVAAS) developed by William Sanders and his associates. Sanders and Horn (1998) describe their model and Baker & Dengke (1995) and Bock, Wolfe, Wolfe, & Fisher, (1996) critique it as implemented in Tennessee.

The Tennessee model appears to be based entirely on gain scores. Other models incorporate gain scores along with other information on students and schools to arrive at school ratings. Clotfelter & Ladd (1996); Heistad & Spicuzza (2000); and Webster, Mendro & Almaguer (1993) describe several of these models.

So far, the major reported use of gain scores has been in rating schools or teachers, particularly by Sanders and Horn (1998). Stone (2002) used the Tennessee system for rating teachers based on student gains to examine whether teachers achieving national accreditation had better achievement among their students.

Gain scores are often used in individual studies and experiments, notably in class size reduction studies, including the Tennessee STAR study (Finn, 1998) and studies of SAGE in Wisconsin (Molnar, Smith, Zahorik, Halbach, Ehrle, Hoffman, & Cross, 2001). Generally these studies have implemented their own testing, rather than using already-existing district or state data. Another example is the growing body of literature concerning the effect of vouchers on student achievement. In several of these programs, a lottery is used to select the participants when too many students apply. Subsequent test scores are used to compare the students selected with those who were not, under the assumption that the only difference was participation in the program.

Surprisingly few districts have used their own data on student gains to analyze educational issues. Issues in which gain scores could be helpful include:

- Whether K-8 schools or middle schools are more effective and for which students.

- When elementary schools are judged effective based on average student scores, how well their students are able to maintain their advantage when they move on to middle school.
- The typical variation of individual student scores from one year to another.
- Whether gain scores pinpoint opportunities for intervention.
- Looking at whether particular programs or strategies have a disproportionate effect on certain kinds of students.
- The effectiveness of particular educational and curricular proposals.
- As a check on data integrity, since a school that artificially inflates test scores in one year is likely to see a decline in later years or when students move to another school.

This study focuses on one issue: what happens to students when they move to high school in ninth grade. I examine why ninth grade seems to be a lost year for most students. I hope this work will encourage other researchers to look at data from other districts. Do they find the same pattern, or is it unique to the district studied here?

Method

Overview of Test Results

For this study, I obtained permission to analyze student achievement data from the Milwaukee Public Schools (MPS), an urban district of about 100,000 students. Roughly 17% of its students

are white, 61% African American, and 15% Hispanic. Seventy seven percent of the students qualify for free or reduced lunch.

For a number of years, all schools in Wisconsin have been required to test students in the fourth, eighth, and tenth grades, in reading, language arts, mathematics, science, and social studies. Wisconsin has adopted the Terra Nova test from McGraw-Hill/CTB.

Starting with the 2000-01 school year, MPS added the Terra Nova reading, language arts, and mathematics tests in grades not covered by the state tests, through tenth grade. Initially the district began testing in the second grade. Beginning with the 2001-02 school year, the district's school board eliminated the Terra Nova tests before the fourth grade except for the mathematics test in third grade.

An important consideration influencing the choice of the Terra Nova was its compatibility with the state tests. As with other standardized tests, the Terra Nova uses item response theory in its construction. As a result, its "scale scores" are comparable from one grade to the next. Thus, a score of 630 on the fourth grade Wisconsin test is equivalent to a score of 630 on the fifth grade MPS test.

Because of score comparability, it is possible to calculate gain scores for individual students. If a student receives a 630 on the fourth grade test and a 640 on the fifth grade test, the gain would be 10 points.[1]

I received the data in the form of separate files for each test, as well as files for each year's enrollment giving information on the student's school, grade, gender, ethnic code, eligibility for subsidized lunch, English language proficiency, and exceptional education status. To reduce the danger of inadvertent release of individual student records, MPS encoded the student numbers. The encoding was consistent from one file to another, allowing me to connect the records using relational database software.

Results

Table 1 shows the average scale scores for all students taking the tests in the spring of 2001 and the spring of 2002. The differences represent the average score in that grade minus the average

Table 1. Average Test Scores by Grade

Average Test Scores									
	Reading			**Language Arts**			**Math**		
Grade	**2000-01**	**2001-02**	**Difference**	**2000-01**	**2001-02**	**Difference**	**2000-01**	**2001-02**	**Difference**
2	588			591			548		
3	615		28	614		22	587	594	39
4	632	631	17	631	636	17	613	616	24
5	646	653	18	645	650	14	626	630	13
6	637	653	-5	643	650	-1	632	634	5
7	649	651	5	643	653	2	646	650	15
8	664	661	13	656	661	11	671	663	19
9	664	658	-1	664	655	1	667	660	-3
10	692	677	23	681	678	20	694	691	29

score in the previous grade. The differences should not be confused with the gain scores discussed later. These differences are calculated using students in different grades tested in the same year. Gain scores, by contrast, are calculated using the same students in different years.

Average reading and language arts scores generally rise each year through fifth grade. They rise again in eighth grade and in tenth grade. Mathematics scores rise fairly steadily through eighth grade, with another sharp rise in tenth grade.

The standard deviations of the scores shown in Table 1 range from 34 to 57 and the test-taking population at each grade varies from 4,400 to 7,000 in the 2001-02 school year. Thus, depending on the grade level and subject, changes in average score between one and a half to two and a half points would be statistically significant at the 95% level.

Table 2 shows the average annual gain for students who took the tests in both the 2000-01 and

Table 2. Average Annual Gains (2000-01 to 2001-02)

Average Gains in Test Scores			
Grade	**Reading**	**Language Arts**	**Math**
2-3			47
3-4	17	23	31
4-5	23	22	19
5-6	11	9	12
6-7	15	11	19
7-8	11	12	17
8-9	-4	1	-7
9-10	7	7	16

2001-02 school years. To calculate these gains, I took a student's 2001-02 score, subtracted that student's score for the previous grade in 2000-01, and then averaged the results for each grade.

The standard deviations of the gains are typically around thirty and the population sizes run from 3,000 to 6,000. At a 95% confidence level, differences in gains more than one or two points would be statistically significant.

To be included in the gain calculation, a student must have taken the test in both years, biasing the results toward the better students. Among the students missing from the calculation are those retained in a grade, those exempted from the tests because of limited English or a learning deficiency, those absent on test day, and those who moved into or out of district schools. In addition, it is often claimed within the district that some schools encourage students expected to depress scores to skip school on test day.

Comparing Table 1 with Table 3 gives a sense of the magnitude of this selection effect. While Table 1 shows the average scores for all students taking the tests, Table 3 shows the averages for only those students taking two tests in sequence. In most cases those taking both tests averaged a few points above the average for all students. The last column in Table 3 compares the size of the population of students taking both tests to the population of all students taking a test.

Table 3. Average scores for students taking tests in both 2000-01 and 2001-02

	Reading		Language Arts		Math			
Grade	**2000-01**	**2001-02**	**2000-01**	**2001-02**	**2000-01**	**2001-02**	**Count**	**Percent**
2	592		586		545			
3	616		615		587	597	5,681	80%
4	644	633	631	638	614	618	5,870	86%
5	646	656	645	653	626	633	5,852	84%
6	640	658	646	654	635	638	5,178	74%
7	654	655	658	658	658	655	4,550	70%
8	664	665	657	668	670	678	4,914	83%
9	674	661	675	660	682	663	4,704	77%
10		682		683		686	3,336	76%

A comparison of Tables 1 and 2 indicates that, before high school, gains exceed the increase in average scores. In fact, if all students were to make these gains starting from third grade (second in mathematics), future eighth grade scores would be thirty points higher than present scores, as shown in Table 4.

Table 4. Actual 8th Grade Scores vs. Predicted Scores

	Reading	**Language Arts**	**Math**
Actual Average	661	661	663
Predicted Average	691	690	692

There are several possible explanations for this paradox. The most optimistic is that present gains are higher than past gains, eventually resulting in higher average test scores in the upper grades.

A less optimistic explanation is that the gains reflect the somewhat selective nature of the student population that consistently takes exams, and that average scores will continue to be depressed by the lower scores of the more erratic test-takers.

Results for Ninth Grade

Both the average scores (Table 1) and the gains (Table 2) show scores holding steady or declining in ninth grade, when practically all students start a new school.[2] To what extent does the apparent ninth grade slump represent an actual pause in student achievement growth, rather than an artifact of testing?

In the past, several observers noted the lack of growth in average scores between eighth and ninth grade. Many, however, chalked it up to a ninth grade enrollment bulge caused by some students' difficulty in earning sufficient credits for promotion to tenth grade. The large number of ninth grade students in academic trouble was assumed to artificially depress the average ninth grade score. But this phenomenon does not explain the pause in gain scores which are calculated based on individual students.

If either the beginning or ending scores have been artificially inflated or depressed the reported gains may be larger or smaller than the actual gains. There is some incentive to inflate eighth grade scores. In Wisconsin the press and public often judge the quality of middle schools by the percentage of their students in each eighth-grade proficiency level. Students from some middle

schools lost more points in ninth grade than did students from other schools, possibly reflecting strategies to exaggerate the score.

Students, in contrast to schools, have no incentives other than pride to do well on the tests. It is likely that for some students a low score reflects a decision not to take the test seriously, rather than underlying academic problems.

How Broad is the Slump?

Separating changes due to variation in student learning from those due to testing problems is difficult. However, very large changes may be particularly suspect as reflecting problems with either the first or second test, rather than changes in learning.

For each of the three tests the standard deviation of the gains is around thirty scale points. Using the common rule of thumb that data beyond three standard deviations can be considered outliers, about 2.2% of the gains might be treated as outliers. This contrasts with the 0.27% of the data that would be expected if gains followed a normal distribution. The relatively large proportion beyond three standard deviations reinforces the suspicion that some of the data may be unreliable. Could the apparent grade slump be an artifact of changes at the fringes?

To better assess whether the slump cuts across various groups, I re-analyzed the data by eliminating students with especially large gains or losses. Table 5 summarizes some of the results. Line 1 shows the average gains for students taking the ninth grade test in the spring of

Table 5. Effect of Filtering on Gains for Various Groups

	Reading Gain	Language Gain	Math Gain	Count
1. Class of 2005: All Students	-3	2	-7	4,704
2. Changes < 3 Sigma	-2	4	-5	4,441
3. Changes < 30 points	-1	4	-1	2,418
4. Blacks with free/reduced lunch	-4	1	-9	2,006
5. Whites w/o free/reduced lunch	-1	8	-1	528
6. Class of 2004: All Students	-1	0	-3	4,529

2002. Line 2 shows the results of eliminating students showing either an increase or decrease greater than three standard deviations (roughly ninety points) in any of their scores between eighth and ninth grade. While there is some improvement in average gains and losses, the effect is slight.

For line 3 any student showing a gain or loss more than thirty points on any test was eliminated, based on the assumption that a one-year gain or loss that large may be suspect. As shown by the count, this filter eliminated almost half the students taking both tests. Still, the effect on the average gain was small.

To further investigate whether the ninth grade slump was concentrated in one or another group of students, I calculated the gains for African American students qualifying for free or reduced lunch (usually one of the groups of students showing the lowest results) and for white students who did not qualify for subsidized lunch. Again the average ninth grade reading and mathematics gains were flat or negative, as shown in lines 4 and 5.

Line 6 in Table 5 shows the gains calculated using the previous class of students (class of 2004). Again, the average ninth grade gains are around zero. As shown in Table 2, however, this same group of students resumed gaining the following year when they took the tenth grade test.

Thus, the overall conclusion is unchanged. On average, students make negligible progress between eighth and ninth grade, at least as measured by the tests. Evidence for the ninth grade slump seems robust: it affects all groups of students and shows up in two different years.

The Effect of Student Achievement

One additional issue is whether the gain or loss is related to the student's initial score. Wright, Horn, and Sanders (1997) in their Tennessee studies found that students with the highest achievement generally had the smallest gains. They suggest possible causes including lack of opportunity for these students to proceed at their own pace, a lack of challenging materials, a lack of accelerated courses, and concentration on average or below-average students.

By contrast, my data shows the ninth grade slump effecting low-achieving students slightly more on average than high-achieving students. Figure 1 shows a scatter plot of the gain in reading from eighth to ninth grade versus the average of each student's reading scores in the two grades. The solid line is the linear trend line which has a slope of .16. The language arts and mathematics tests also show slightly stronger relationships, with slopes of .23 and .24 respectively. Because of the large number of students, all these results are statistically significant at any reasonable

confidence level. (I eliminated students achieving either the maximum or the minimum scores from this calculation.)

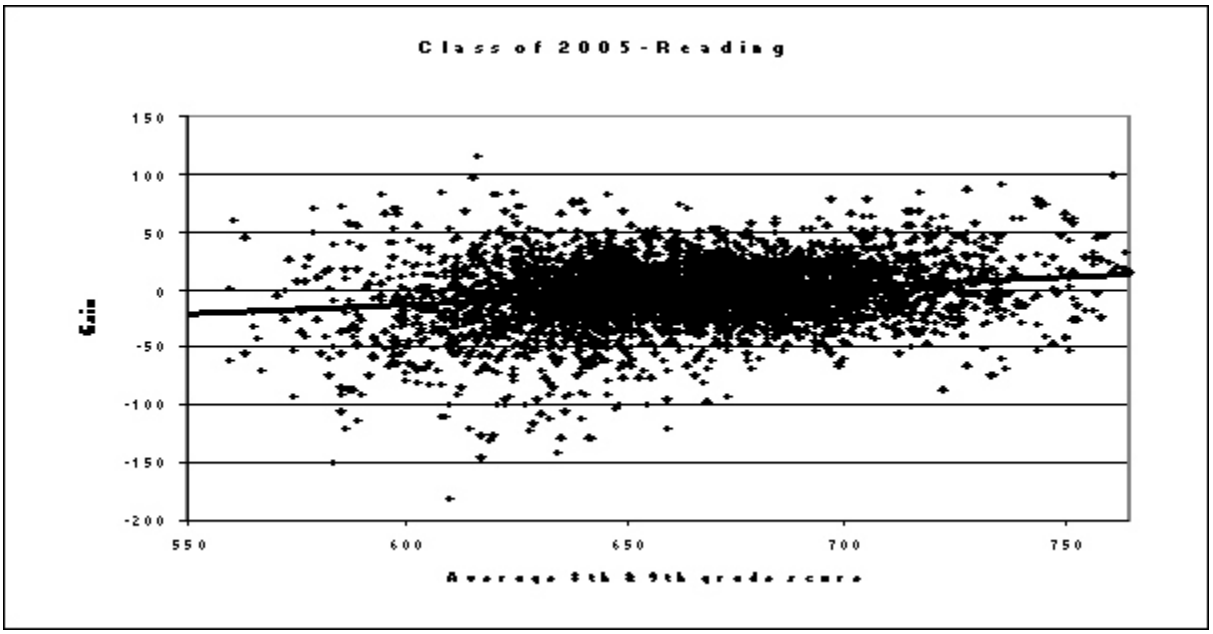

Figure 1. Comparison of reading gain to average score

Figure 1 is consistent with the result shown in Table 5, where the ninth grade slump seems to affect middle-class white students slightly less than low-income black students. Sanders and Rivers (1996) concluded that minorities and low-socioeconomic students were more likely to be assigned to less effective teachers whose students make much lower achievement gains than students with more effective teachers.

Discussion

As seen in Table 3, something happens when students move from eighth grade to ninth grade. On average, their scale scores stay constant or decline. In addition, the relationship between average

gains and the differences of the averages reverses itself. The gains become less than the differences between the average score in one grade and that in the following grade. Based on the gain scores, it appears the average student learns nothing in ninth grade, at least of the material included on the tests.[3]

Sanders, Saxton, Schneider, Deardon, Wright, & Horn (1994) reported a loss of expected gain whenever students transfer to the lowest grade in a building. Interestingly, they did not find a loss when students transferred to subsequent grades. They associated this loss with the move from elementary school to middle school. The Tennessee assessment program extends only through eighth grade, so they were unable to look at the transition to high school.

Sanders, et. al. (1994) suggested three possible factors contributing to the lower gain:

1. A breakdown in communication between sending and receiving schools leading to excessive reteaching and lack of continuity.
2. The acclimation and processing of entering students cutting into time for instruction, and
3. A loss of teaching time in the first weeks of schools while teachers become acquainted with their students' abilities and achievement levels.

It seems likely that the same factors contribute to the ninth-grade slump. If high schools underestimate the achievement levels of their incoming freshmen and assume that students come

out of middle school knowing little they may concentrate on bringing their incoming students up to a satisfactory level, in essence repeating eighth grade for the average student.

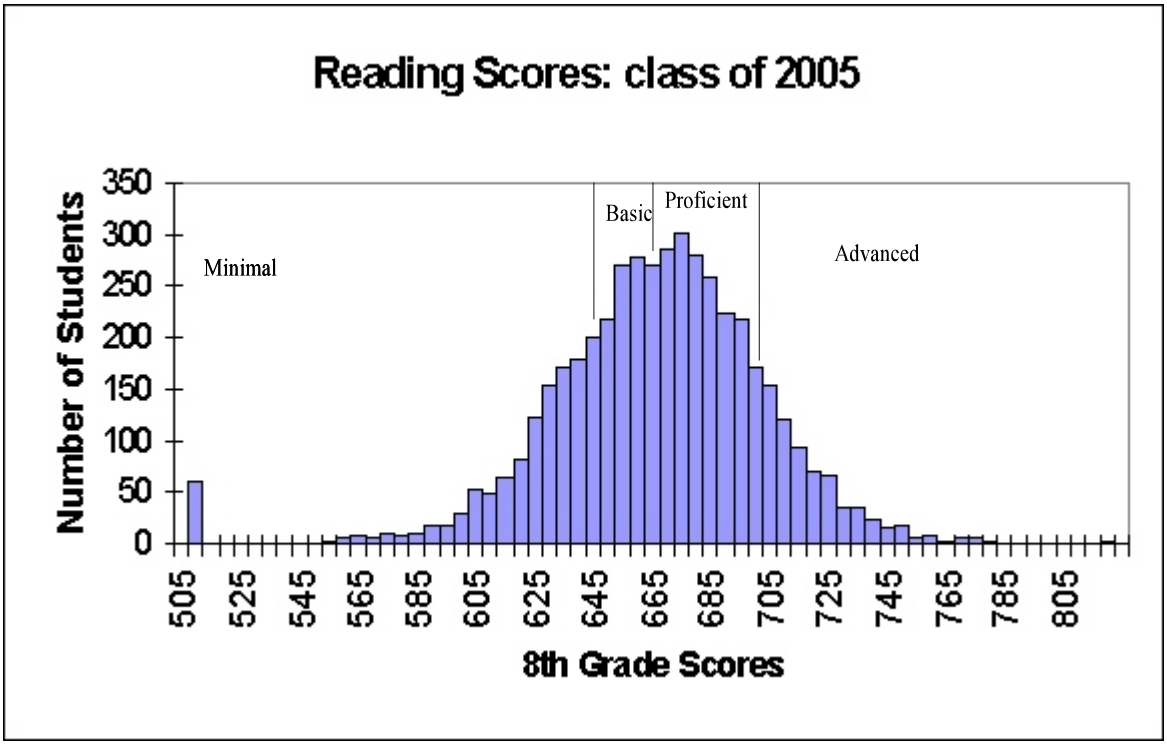

Figure 2. Distribution of eighth grade reading scores

But to speak of average students may be misleading. Urban high schools are confronted with a freshman class whose preparation varies widely. Figure 2 shows the distribution of eighth grade reading scores for the incoming high school freshmen.

Consider the distribution of reading abilities a typical freshman high school teacher may confront in one class of thirty students, as shown in Table 6. If the class is typical of the school system as a

Table 6. High School Class of 30: State proficiencies

	System Average	High Input School	Low Input School
Minimal	11	5	20
Basic	6	8	6
Proficient	11	12	4
Advanced	2	5	0
Total	30	30	30

whole, eleven of the students will have been judged by the state to be reading below the minimal level for eighth grade while thirteen will be proficient or advanced.

In Table 7, members of this class are compared to the average scores of student in various grades. Five of the students read below the school system's average fourth grader while eleven already exceed the average tenth grader's reading level.

Table 7. High School Class of 30: Average reading levels

	System Average	High Input School	Low Input School
Below 4th	4	2	11
4th to 6th	6	2	8
6th to 8th	3	1	3
8th to 10th	6	4	5
Above 10th	11	22	3

Some schools, of course, differ from the system average but their task may be no easier. At the high school with the highest incoming test scores, four students out of thirty would be below basic in reading and two below the average fourth grader. At the school with the lowest-scoring incoming class, twenty of the students would be below basic. Yet another four would be rated proficient.

After ninth grade, as schools and teachers learn more about their students, they may use strategies to challenge them at a level more appropriate to their abilities. In later grades high schools sort their students more. For the better-prepared students, there are honors programs, advanced placement courses, and International Baccalaureate programs. By tenth grade students are learning once again, as reflected in the last line of Table 5.

This hypothesis suggests a strategy to combat the high school slump:

1. Information on incoming student abilities and achievement collected by both the middle schools and the school system should precede the students' first day in high school. As districts collect more data on students and their scores, schools could use that information on incoming students to make incoming class assignments and to design intervention programs for those students falling behind.

2. Make sure the high schools have courses that meet the very wide range of student needs. These would range from intensive remedial help in reading and mathematics to preparation for Advanced Placement and International Baccalaureate courses.

3. Use the data to help students choose the course level most likely to challenge them without overwhelming them.

Currently, much of the concern over urban high schools focuses on the problem of the over-challenged student, who lacks the preparation needed to succeed, gets discouraged, and drops out. Finding ways to help those students gain the skills they lack is certainly important.

The present study suggests that the majority of high school students may be in the opposite category: under-challenged students who could have learned much more in the ninth grade. In many cases, these students get good grades, take their schools' most challenging courses, and yet find themselves behind their suburban counterparts when they enter college. Since the majority of minority students come from urban school systems, the problem of under-challengement may help to explain minority under-representation at the highest levels of education.

References

Baker, A. Paige; Xu, Dengke (1995) The Measure of Education: A Review of the Tennessee Value Added Assessment System. Nashville: Tennessee State Comptroller of the Treasury,. Office of Educational Accountability.

Bock, R. D., Wolfe, R. D., Wolfe, R., & Fisher, T. H. (1996). A review and analysis of the Tennessee Value-Added Assessment System. Nashville, TN: Tennessee Controller of the Treasury.

Clotfelter, C., & Ladd, H. (1996). Recognizing and rewarding success in public schools. In H Ladd, Holding schools accountable, Chapter 2. Washington, DC: The Brookings Institution.

Cooney, S., & Bottoms, G. (2002). Middle Grades to High School: Mending a Weak Link. Atlanta: Southern Regional Education Board . Available at http://www.sreb.org/programs/hstw/publications/briefs/MiddleGradestoHS.asp

Finn, J. D., (1998) Class size and students at risk: What is known? What is next? Washington, DC: U.S. Department of Education.

Heistad, D., and Spicuzza, R. (2000). Measuring school performance to improve achievement and to reward effective programs. Presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA, April 2000

Meyer, R.H. (1997). Value-Added Indicators of School Performance: A Primer, Economics of Education Review. 16:3, pp 283-301.

Molnar, A., Smith, A., Zahorik, J., Halbach, A., Ehrle, K., Hoffman, L., & Cross, B. (2001) 2000-2001 Evaluation Results of The Student Achievement Guarantee In Education (Sage) Program. Milwaukee, WI: School of Education, University of Wisconsin—Milwaukee. *Available:* http://www.uwm.edu/Dept/CERAI/sage.html

Sanders, W.L. and Horn, S.P. (1998). Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research, Journal of Personal Evaluation in Education. 12:3, 247-256.

Sanders, W.L. and Rivers, J.C. (1996) Cumulative and Residual Effects of Teachers on Future Student Academic Achievement. Knoxville TN: U of Tennessee Value-Added Research and Assessment Center. Available at http://www.mdk12.org/practices/ensure/tva/.

Sanders, W.L., Saxton, A.M., Schneider, J.F., Deardon, B.L., Wright, S.P., & Horn, S.P. (1994). Effects of building change on indicators of student academic growth. Evaluation Perspectives, 4(l). Available at http://www.mdk12.org/practices/ensure/tva/.

Stone, J. E. (2002), The Value-Added Achievement Gains of NBPTS-Certified Teachers in Tennessee: A Brief Report, College of Education, East Tennessee State University. Available at http://www.education-consumers.com/briefs/stoneNBPTS.shtm

Thompson, Bruce R. (In press). Equitable Measurement of School Effectiveness, Urban Education.

Webster, W., Mendro, R., & Almaguer, T. (1993). Effectiveness indices: the major component of an equitable accountability system. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

White, S.B., Reynolds, P., Thomas, M., & Gitzlaff, N. (1993). Socioeconomic Status and achievement revisited, Urban Education 28, 328.

Wright, S.P., Horn, S.P., & Sanders, W.L. (1997) Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation, Journal of Personnel

Evaluation in Education, 11(1), 57-67. Available at http://www.mdk12.org/practices/ensure/tva/.

Endnotes

[1] While a higher score indicates a student has answered questions of greater difficulty, there is no assurance that the scales are linear. Thus there is no easy way to compare growth at one level to growth at other levels. For example if one student progresses from a score of 500 to a score of 510, that student has gained 10 points. A student going from 630 to 640 has also gained 10 points. But from this information one cannot conclude that the two students have made the same progress or, if not, which student gained more.

[2] Another transition year, that from fifth to sixth grade, also shows a decline in average reading and language arts scores. The gains, however, are still positive but lower than in either of the immediately adjacent years. Analysis of this transition is complicated by several factors, including possible inflated results at a few schools and by the trend toward K-8 schools.

[3] The slump may not be unique to incoming high school freshmen. There is some evidence in the present data of a pause in gains among students who move from elementary schools to middle. The evidence is more ambiguous because it seems to vary by subject, their seem to be exaggerated scores coming from some elementaries, and some students stay in the same school.