WCER Working Paper No. 2003-5
April 2003

Value-Added Indicators: Do They Make an Important Difference? Evidence From the Milwaukee Public Schools

Robert H. Meyer
Wisconsin Center for Education Research
University of Wisconsin–Madison
rhmeyer@aol.com

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison
http://www.wcer.wisc.edu/

The research reported in this paper was supported by the Joyce and Helen Bader Foundations and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the author and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Value-Added Indicators: Do They Make an Important Difference? Evidence From the Milwaukee Public Schools[1]

Robert H. Meyer

Value-added indicators are increasingly being used throughout the nation to provide valid evidence of the performance (or productivity) of schools, programs, and education policies. The move toward research-based methods of evaluation has been motivated by several factors:

- Greater focus on academic outcomes as measured by achievement tests, such as the TerraNova
- Recognition that some schools are doing a much better job than others at producing gains in achievement, particularly among disadvantaged students
- Greater availability of achievement data

The Milwaukee Public Schools initiated the annual assessment of students in 2001. The availability of this annual assessment data puts the district in a position to construct value-added indicators to track the performance of schools, programs, and policies to improve the academic performance of students. The first part of this paper uses mathematics achievement data from Milwaukee seventh- and eighth-grade students to explain and illustrate the power of the value-added approach. The second part of the paper explains why value-added measures of school performance are superior to attainment (non-evaluation-based) indicators, such as average test scores.

Value-Added Indicators

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. The indicators are derived from a statistical model that includes, to the extent possible, all of the non-school factors that contribute to *growth in student achievement* — in particular, prior student achievement and student, family, and neighborhood characteristics. The purpose 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. The end result is a value-added indicator that captures differences in educational productivity among schools, programs, and policies.

The logic of the value-added approach is illustrated by Figure 1, which compares test-score information for students from two Milwaukee middle schools with code names A and LL (the same schools identified by these code names in Table 2). Each point on the graph represents test score data for a single student, with eighth-grade mathematics achievement on the vertical

[1] Presented at the annual meeting of the American Educational Research Association, New Orleans, April 2, 2002. The author thanks Bridget Gleeson for helpful research assistance, Bill Clune and Sarah Mason for comments on the paper, and Cathy Loeb and Margaret Powell for editorial assistance.

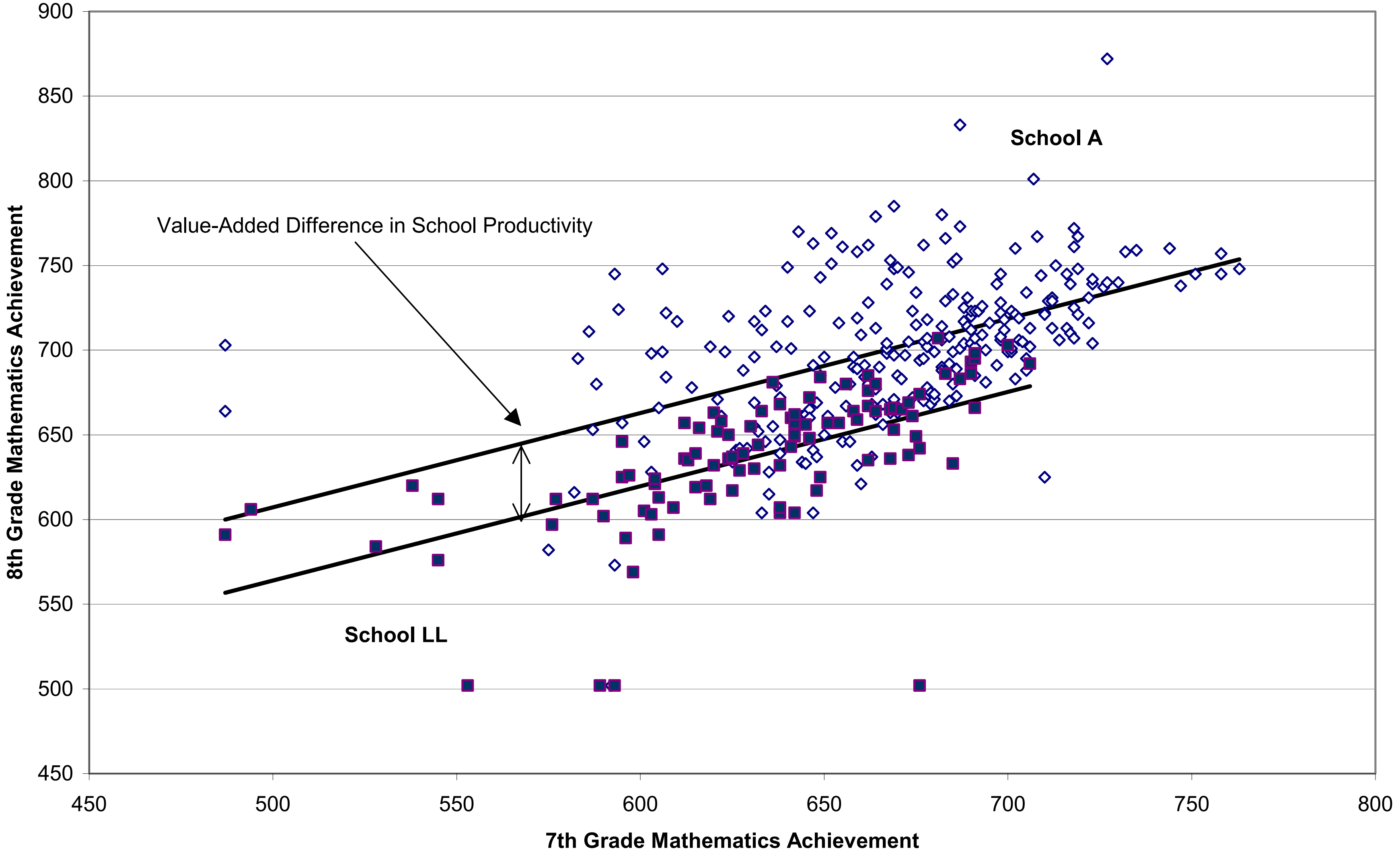
axis and seventh-grade mathematics achievement on the horizontal axis. The data points for students in School A are represented by diamonds. The data points for students in School LL are represented by squares. As one would expect, students with higher seventh-grade test scores tend to have higher eighth-grade test scores. On the other hand, this relationship is far from exact. Eighth-grade test scores vary significantly, even for students with the same seventh-grade mathematics scores. This variation is due to differences across students in the non-school-related factors that influence student achievement and to errors (noise) in measuring student achievement. Despite the variation in the data, it is evident that there is a systematic difference in eighth-grade achievement between the two schools. For comparable students (that is, students with the same seventh-grade achievement), eighth-grade achievement is substantially higher, on average, for students in School A than for students in School LL. The average achievement levels for both schools are represented on the graph by the two lines (regression lines).[2]

The distance between the two regression lines captures two components: (a) the value-added difference in the productivity of the two schools and (b) a possible difference between the schools in non-school factors related to growth in student achievement (for example, differences in student mobility). If we assume that the latter component is small, then the difference in the two regression lines provides an estimate of the difference in the value-added productivity of schools A and LL (43.2 scale-score points). In general, we can construct more refined (accurate) value-added indicators by expanding the analysis to include other "control" variables in addition to seventh-grade mathematics achievement—for example, student characteristics such as student mobility and demographic factors. A value-added model that controls for multiple variables is more difficult to illustrate graphically, but it is no more difficult to implement than a model with only a single control variable (seventh-grade mathematics achievement). Moreover, the logic of the two types of models is very similar. As a result, in the next section I use a value-added model with multiple controls to demonstrate the value-added approach to measuring middle school performance in Milwaukee. As indicated in Table 2, the estimated productivity difference between Schools A and LL based on the more complete value-added model is equal to 28.1 scale-score points, somewhat less than the estimate obtained from the simple model, but nonetheless very large (more on this below).

In summary, the objective of a value-added analysis is to develop a model of student achievement that provides the best possible estimates of the productivity of schools (programs and policies). See Meyer (1996, 2000) for additional information on the value-added approach.

[2] The value-added approach does not require the lines in Figure 1 to be linear or parallel. It is conventional to test for whether these restrictions (which greatly simplify the analysis) are acceptable, given the data.

Figure 1. A Graph of Student Achievement Data for Two Schools



A Value-Added Model of Mathematics Achievement Growth From Seventh to Eighth Grade in Milwaukee

In the empirical results reported below, value-added indicators were obtained from a model of eighth-grade mathematics achievement (measured in February 2000) that included the following control variables: seventh-grade mathematics achievement (February 1999); gender; race/ethnicity; an indicator of economic disadvantage (participation in the free or reduced-price lunch program); and an indicator of student mobility from 2000 to 2001.[3] Achievement scores were measured using the TerraNova assessment and reported as a developmental scale score. The sample used in the analysis was a matched longitudinal sample; it was restricted to students who had seventh- and eighth-grade mathematics scores in 2000 and 2001, respectively.[4,5] The matched sample included 4,751 students. In contrast, the sample of students with eighth-grade mathematics scores (and possibly missing seventh-grade scores) included 5,603 students.

The matched sample consisted of a group of students with the following characteristics (see Table 1):

- 51% female
- 63% African American
- 20% White
- 11% Hispanic
- 6% Native American, Asian, or other race/ethnicity
- 73% free or reduced-price lunch
- 10% mobility (changed schools between February 1999 and February 2000)

[3] The value-added model includes the following refinements, all of which enhanced the accuracy and precision of the estimates: One, the model corrects for measurement error in the achievement measures. (See Meyer, 1999, for a discussion of methods designed to correct for measurement error.) Two, to maximize the precision of the estimates and to be sure that all students were represented, students who changed schools between February 2000 and February 2001 were included in the analysis sample. Note also that the reported estimates of school performance pertain to the period between the February test dates—that is, the spring of seventh grade and the fall and winter of eighth grade. The estimates thus reflect the partial contributions of seventh- and eighth-grade educators from the same school to growth over this period. (As an aside, I should note that testing in February makes the analysis a little trickier than if students were tested in May or September.) Three, the model was estimated using multilevel methods.

[4] No students were eliminated due to missing data on the other variables in the model.

[5] This restriction inevitably eliminates students who move into a district after the testing window or who were absent during testing. One way to include in-migrant students in the analysis is to implement a point-of-entry testing system that assesses all such students at the time that they enter the district. Minneapolis is experimenting with such a system.

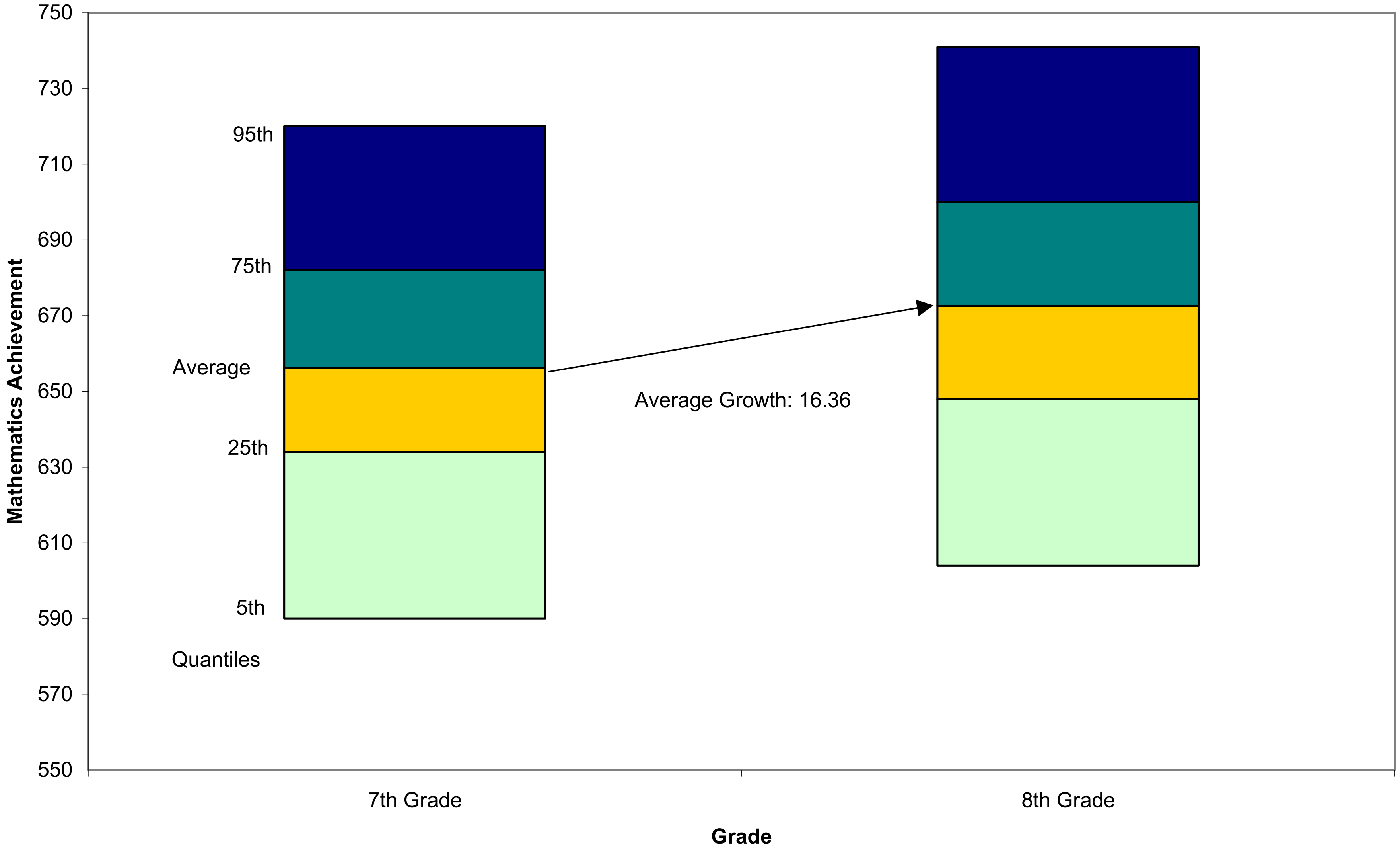
These sample characteristics closely match statistics reported by the district for all eighth-grade students in school year 1999–2000.

Table 1
Simple Statistics

Variable	Mean (Standard deviation of non-binary variables)
Eighth-grade mathematics achievement	672.58 (44.52)
Seventh-grade mathematics achievement	656.22 (41.54)
Female	0.51
Native American	0.01
African American	0.63
Asian	0.040
Hispanic	0.11
Other	0.01
White	0.20
Free lunch	0.62
Reduced-price lunch	0.11
Mobility: changed schools	0.10
Sample size	4751

As indicated in Table 1 and Figure 2, average mathematics achievement increased from a scale-score value of 656.22 in seventh grade to 672.58 in eighth grade, an average increase of 16.36 scale-score points. This increase, although substantial, was small relative to the variability in test scores in seventh or eighth grade. In particular, the standard deviation of test scores was quite a bit larger in both seventh and eighth grade: 41.54 in seventh grade and 44.52 in eighth grade. Similarly, the range of test scores from the 5th to the 95th quantile was much larger: 130 points in seventh grade and 137 points in eighth grade. This pattern is typical: the growth in test scores in any one year tends to be small relative to the spread in test scores that accumulates over multiple years of learning. There is also significant variability in growth in test scores in a given year. This phenomenon was observed earlier in the graph of seventh- and eighth-grade test scores in Figure 1.

Figure 2. The Average and Range of Test Scores in 7th and 8th Grade



Next, I consider estimates of the value-added performance of Milwaukee middle schools from February 1999 (seventh grade) to February 2000 (eighth grade). This period captures the combined contributions of spring of seventh grade and fall and early winter of eighth grade.[6] Table 2 reports estimates of value-added school performance and, for reference purposes, average eighth-grade mathematics achievement. In this table, the primary value-added indicator is reported in a form referred to as a "beat the average" indicator. This indicator should be interpreted as measuring school performance relative to the district's average performance across all schools in seventh and eighth grades. Hence, a school with a value-added value of 0 would be considered an average quality school in the district (with respect to mathematics achievement) at this grade level.[7] Table 2 also reports a related indicator, the value-added performance tier. This indicator reports performance on a normative scale that generally ranges from 1 to 5, where 3 indicates an average school, 5 indicates a top-tier school, and 1 indicates a bottom-tier school.[8] The advantage of the performance tier indicator is that it can be readily interpreted without reference to the units of the achievement test. The advantage of the "beat the average" value-added indicator (and similar indicators) is the opposite: it is defined in terms of the units of the achievement test. The two indicators provide somewhat different information and thus are both useful. The value-added estimates reported in Table 2 are sorted by estimated performance (from top to bottom). School names are not included in the table. Instead, I have identified the schools by an alphabetic code that ranges from A to PP.

What does Table 2 tell us about middle schools at the seventh- and eighth-grade level? There is good news and bad news. First, the good news: The top-tier middle schools in Milwaukee did a great job of generating growth in mathematics achievement in 1999–2000. The top school (School A) produced growth in test scores equal to 15.0 scale points compared to the average school. To get an idea of the significance of this effect, note that achievement growth for the average student in the average Milwaukee middle school was equal to 16.4 scale points. In other words, the average student in School A experienced almost double the achievement growth (16.4 + 15.0 = 31.4 scale points) compared to the average student in the average Milwaukee middle school. Now, the bad news: The bottom-tier schools performed quite poorly, receiving "beat the average" values of less than –12.0 scale points, equivalent to essentially no achievement growth. In addition, two schools performed substantially worse than other schools and fell into performance tiers 0 and –1. On average, students in these schools had lower mathematics scores in eighth grade than in seventh grade.

[6] One major advantage of testing students in early May (as is done in many school districts) is that the period between testing includes an entire school year, rather than parts of two school years.

[7] In order to compare two schools, one simply subtracts one indicator from the other.

[8] The performance tier is defined as $P = 3 + (\hat{\alpha} - \overline{\alpha}) / \sigma_{\alpha}$, where $\hat{\alpha}$ = estimated school performance, $\overline{\alpha}$ = average school performance at some point in time (zero, by definition in the "beat the odds" form of the value-added indicator), and σ_{α} = the (noise-corrected) standard deviation of school performance. This statistic is similar to a standard z-statistic except that it is centered on 3, not 0, and it is rounded to the nearest integer. More generally, the statistic could be rounded to a single decimal place to preserve greater information when comparing schools. Very high-performing schools could have performance tier values in excess of 5. Similarly, very low-performing schools could have performance tier values less than 1.

Table 2
School Performance and Average Eighth-Grade Mathematics Achievement

School	Performance Tier	School Performance: Beat the Average	Average 8th Grade Math Achievement
District	3	0.0	670.0
A	5	15.0	700.5
B	5	14.7	634.7
C	5	12.8	692.1
D	5	12.6	703.7
		12.0	
E	4	11.4	700.9
F	4	11.0	655.8
G	4	9.6	639.9
H	4	8.8	702.9
I	4	8.1	663.1
J	4	4.9	684.0
K	4	4.8	696.5
L	4	4.2	658.8
M	4	4.1	660.3
		4.0	
N	3	3.1	681.6
O	3	2.9	686.0
P	3	2.6	689.2
Q	3	2.6	667.4
R	3	1.4	670.6
S	3	1.4	653.8
T	3	0.6	692.3
U	3	0.0	633.5
V	3	-0.1	670.1
W	3	-1.5	664.8
X	3	-1.9	670.7
Y	3	-2.5	662.8
Z	3	-2.8	690.1
AA	3	-2.8	660.5
BB	3	-3.7	662.7
		-4.0	
CC	2	-4.1	658.5
DD	2	-4.5	676.1
EE	2	-4.9	672.6
FF	2	-4.9	682.0
GG	2	-6.6	672.4
HH	2	-7.2	647.5
II	2	-8.4	679.1
JJ	2	-9.9	656.5
KK	2	-11.8	651.7
		-12.0	
LL	1	-13.1	638.8
MM	1	-14.5	659.0
NN	1	-14.6	663.0
OO	0	-20.4	616.8
PP	-1	-31.4	640.1

The bottom line is that in 1999–2000 the productivity of middle schools in Milwaukee varied substantially. Some produced almost no growth (or even negative growth) in mathematics achievement. Others generated almost twice as much growth as the average Milwaukee middle school. As a check on these results, I examined the statistical error of the school performance estimates. Although the error levels were high for very small schools, the reliability of the school performance estimates was very high (89.4%), and the average standard error was relatively small (2.76 scale points).[9,10]

Performance of Schools with Disadvantaged Students

For the last several years, the Milwaukee Public Schools have conducted an analysis that is very much in the spirit of value-added analysis, identifying schools that meet the so-called "90-90-90" criterion:

- 90% or more of their students are minority;
- 90% or more of their students are eligible for free or reduced-price lunch; and
- 90% or more of their students score at or above the basic level on the Wisconsin Reading Comprehension Test (WRCT).

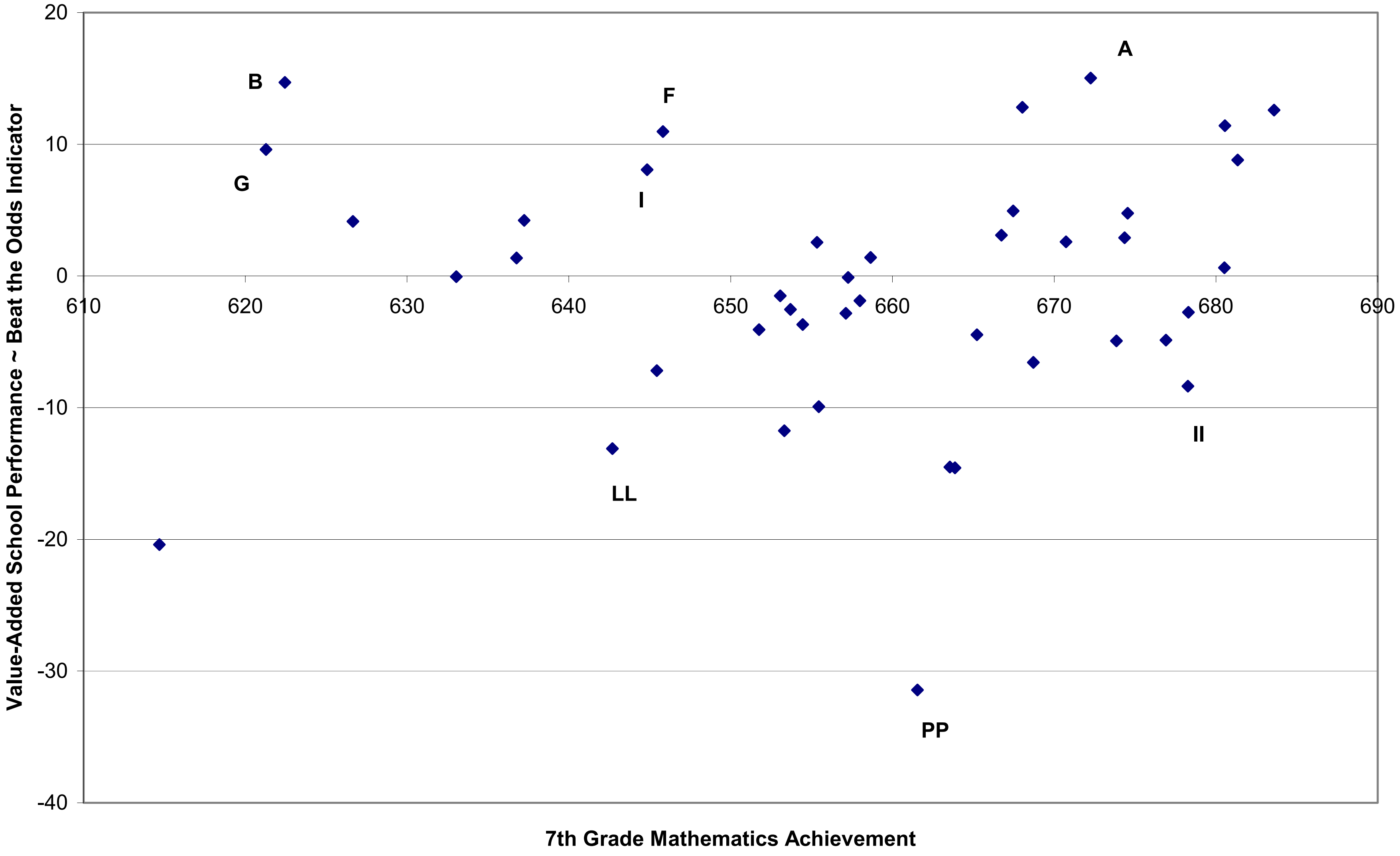
At the heart of the 90-90-90 analysis is a search for schools that are doing an excellent job of educating disadvantaged students. Given value-added estimates of school productivity, one can substantially expand on the 90-90-90 approach. For example, Figure 3 plots school performance estimates by average prior (seventh-grade) achievement. As indicated in the figure, most middle schools with low prior achievement were doing a better than average job of producing growth in mathematics achievement (e.g., Schools B, F, G, and I), and a large number of middle schools with high pretest students had below-average performance (e.g., Schools II and PP).

In the next section, I explain why traditional attainment indicators such as the average test score fail to measure school productivity.

[9] The reliability index measures the share of the variance in estimated effects that is due to true variance as opposed to estimation error.

[10] An important issue, to be addressed in a future paper, is how to improve the precision (statistical accuracy) of school performance estimates for schools with limited enrollment. The simplest and most promising approach is to average school performance estimates over several grades and years. This, in effect, increases the number of students used to estimate school performance.

Figure 3. School Performance by Average Prior Achievement



What's Wrong With the Average Test Score as a Measure of School Performance?

The average test score and other attainment indicators reflect the contributions of schools and other non-school inputs to the learning process from multiple grades and multiple points in time. Such indicators are highly flawed as a measure of school performance for four basic reasons. First, the average test score is *contaminated by factors other than school performance*—in particular, prior student achievement and the average effects of student, family, and community characteristics on student achievement growth. In fact, it is quite likely that comparisons of average test scores across schools primarily reflect these factors rather than genuine differences in intrinsic school performance. As such, average test scores are highly biased against schools that disproportionately serve academically disadvantaged students and communities.

Second, the average test score reflects information about school performance that tends to be *out-of-date*, particularly for later grade levels. Consider, for example, the average test score for a group of students tested at the end of 10th grade. The average test score for this group reflects the accumulated learning that occurred during 10th grade; in 9th grade, 1 year earlier; in 8th grade, 2 years earlier—all the way to kindergarten and preschool, 10 (or more) years earlier. Indeed, a 10th-grade-level indicator could be dominated by information that is 5 or more years old. One consequence of this is that changes over time in average test scores could be completely unrelated to, or even negatively correlated with, actual changes in school performance. Meyer (1996, 2000) reported that this pattern shows up in the NAEP mathematics data: average NAEP scores in 11th grade increased in the late 1980s, following publication of the *Nation at Risk* report (National Commission on Excellence in Education, 1983), despite the fact that the productivity of high schools declined somewhat. Eleventh-grade scores went up because gains in elementary school for this group of students were large enough to outweigh declines in gains during high school. 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.

Third, average test scores at the school, district, and state levels tend to be highly *contaminated due to student mobility*, particularly between schools, but also between districts. For example, the typical high school student is likely to attend several different schools over the period spanning kindergarten through 12th grade. For these students, a test score reflects the contributions of at least two 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.

Fourth, the average test score *fails to localize* (pinpoint) 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. An average 10th-grade test score unfortunately provides almost no information distinguishing the performance of 9th- and 10th-grade teachers. A performance indicator that fails to localize school performance in a specific grade level or classroom is likely to be a relatively weak instrument of public accountability.

One further weakness of the average test score as a measure of school performance is that it creates a pernicious incentive for schools to “cream”—that is, to raise measured performance by educating or testing 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 occur 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 avoid testing, 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 might gravitate to schools and programs that predominantly serve high-scoring students in order to be a part of a school with high, albeit incorrectly measured, school performance.

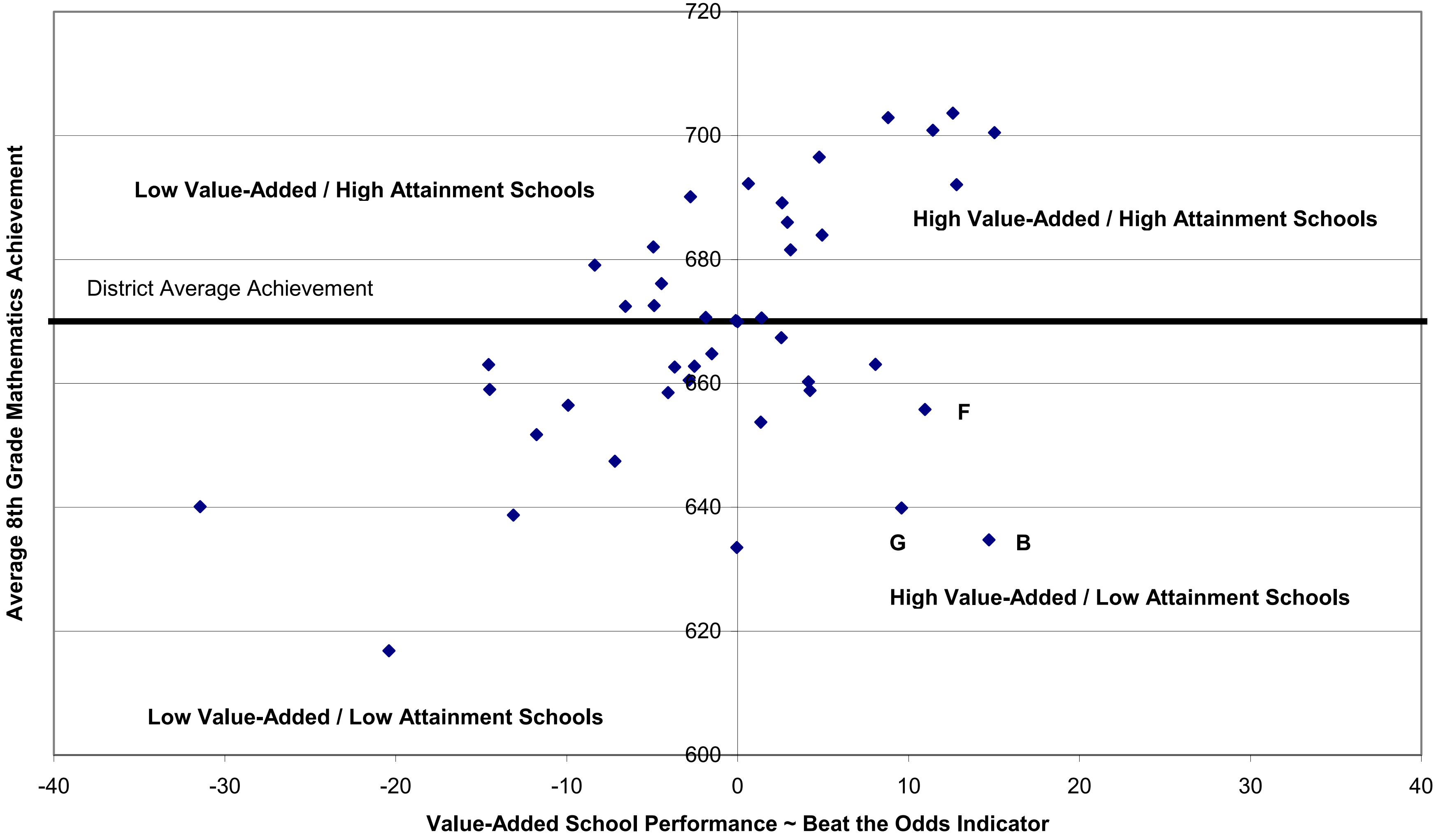
Value-Added Indicators Compared to Attainment Measures

Above, I argued that simple attainment indicators, such as the average test score, were highly questionable as measures of school performance. Here, I investigate whether the average eighth-grade mathematics score and the value-added indicator reported in this paper yield different results. As indicated in Figure 4, although there is a modest positive relationship between the two indicators, there are also major differences. Schools B, F, and G are examples of schools that are rated high with respect to the value-added criterion, but low with respect to average eighth-grade achievement. As was indicated in Figure 3, these three schools all served students with low prior achievement, and the value-added criterion correctly picks up the fact that these schools produced large gains in mathematics achievement from seventh to eighth grade. Overall, eight schools were classified as high value-added but low attainment schools. Similarly, six schools were identified as low value-added but high attainment schools. The value-added criterion correctly picks up the fact that these schools generated very little achievement growth, despite having students with initially high prior achievement. The bottom line is that the value-added approach identifies the up-to-date performance of schools, something that the average test score is not, in general, able to do.

Conclusions

This paper has provided a brief introduction to value-added indicators and their potential use in the Milwaukee Public Schools. Additional test score data that will soon be available in Milwaukee will make it possible to apply the types of analyses demonstrated here to Grades 4 through 10 for the 2001-2002 school year. By 2003 (and in subsequent years), it will be possible to repeat the value-added analyses for these grades, thereby allowing Milwaukee schools (and the district as a whole) to determine whether they are, in fact, improving in their efforts to better educate students, as measured by value-added indicators. That is, it will be possible to evaluate whether value-added improvements increase, stay the same, or decrease over time.

Figure 4. A Comparison of Indicators: Average 8th Grade Mathematics Achievement vs. Value-Added School Performance



References and 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, CA: Academic Press.

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

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: Proceedings of a conference held at the Federal Reserve Bank of Chicago* (pp. 69–97). Chicago: Federal Reserve Bank of Chicago.

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.

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 Institution Press.

Meyer, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief, 3*(3). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Meyer, R. H. (2001). An evaluation of the effectiveness of the Urban Systemic Initiative and other academic reforms in Texas: Statistical models for analyzing large-scale datasets (Report prepared for the National Science Foundation). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research.

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

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Department of Education.

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