EQUITABLE MEASUREMENT OF SCHOOL EFFECTIVENESS

Bruce R. Thompson
Milwaukee School of Engineering

This article examines a value-added model to rate schools, taking into account varying characteristics of students entering the schools, including poverty, race, mobility, and stability. The model uses regression to separate the influence of demographics that are outside the schools' control from those under school control. This value-added model leads to school ratings that compare actual performance to performance predicted by poverty rates. Even after adjusting for poverty, some schools consistently outperform others on a wide range of tests over a 5-year period, whereas other schools consistently perform at the bottom. Finally, this article discusses uses that can be made of this model.

Keywords: *value-added; assessments; testing; standardized testing; school ratings*

Large urban school districts collect and report an increasing variety and quantity of student achievement data, particularly test scores. They also report demographic data on their schools and students.

Controversy has accompanied this expanded supply of data. How should it be used to identify schools and programs that demonstrate unusual student achievement? How can useful information be extracted while practicing equity in assessment?

Newspapers and other observers often turn raw test scores into school rankings (for example, see Borsuk, 2002). As a result, urban schools serving low-income students usually dominate the bottom rankings, whereas schools in wealthy suburbs occupy the upper ranks. This approach leaves the reader in the dark as to how much

AUTHOR'S NOTE: *The author wishes to thank the Lynde and Harry Bradley Foundation, whose grant to the Milwaukee School of Engineering allowed the time to explore the issues described in this article.*

URBAN EDUCATION, Vol. 3X No. X, Month 2004 1-29
DOI: 10.1177/0042085903261325

these rankings reflect school quality differences and how much they reflect differences in the student populations served. If, for example, schools traded student populations, would the suburban schools retain their higher rankings?

Since at least the Coleman report (Coleman et al., 1966), the strong impact of student socioeconomic status (SES) on student achievement has been well accepted. On average, students from low-income families score substantially lower than students from higher-income families on standardized tests and other academic measures.

If family economics are so crucial, how can we fairly compare school performance from one district or school to others? How much can a school affect student achievement? The answers to these questions are crucial to the future of academic reform. If school measurement is to be effective, it must be applied equitably.

At the extremes, two conceptual models dominate the accountability debate. One side, favored by many educators, interprets research like that of Coleman et al. (1966) as showing that the school's influence is minimal. In this view, family, environment, neighborhood, and economics largely swamp whatever the school can do.

An opposing group, including many political and business leaders, holds that schools serving poor children should be held to the same standards as those serving the middle class. If poor children do not do as well, they argue, the school is at fault, reflecting low expectations, a weak teaching staff, poor leadership, or inadequate resources.

The rhetoric in this debate is often heated. Those emphasizing SES are accused of "writing off" low-income students. Those pushing for higher standards are accused of contributing to the high dropout rate among low-income and minority students.

Each position has inherent problems. Fixation on student background can absolve the school of responsibility for student performance At worst, it encourages the familiar urban school fatalism of teachers and principals expecting little from their low-income students.

Ignoring the effects of poverty, however, defies common sense and the results of most research. It implies that the family role in

children's educations is negligible, at least in a good school. By making no adjustments for the disadvantages that low-income students suffer, this approach can exacerbate the difficulty in recruiting and retaining qualified teachers and principals at schools serving the neediest students.

The search for models that separate school and environmental variables has continued over much of the second half of the last century. As early as 1973, Innes and Cormier commented that "attempts to separate out the influence of school variables from community variables have not met with much success" (p. 2), referring to sources dating back to the 1930s. A few school districts have started to develop school performance models that adjust for demographic high-risk factors. Among the better-known models are those used in Minneapolis (Heistad & Spicuzza, 2000), South Carolina (Clotfelter & Ladd, 1996), and Dallas (Webster, Mendro, and Almaguer, 1993). The search for reliable school indicators is international as well (Fitz-Gibbon & Tymms, 2002).

An extensive school effectiveness literature has developed. Teddlie and Reynolds (2000) summarized the major currents of this research. Although the school-effectiveness movement has gathered a network of researchers in several countries, it seems to have had little impact on school practices, at least in the United States. Some critics have complained that school effectiveness research understates the effect of SES or "contextural variables" (see Thrupp, 2001, and the responses in the same journal).

Every model has its critics and defenders. In pursuing scrupulous fairness, some assessment models grow enormously complex and hard to understand (see, for example, Clotfelter & Ladd, 1996, on the Dallas model). Particularly when the model is attached to rewards and punishments, it may become a target of gaming by schools or teachers.

In searching for more accurate models, researchers have taken several approaches. These include:

- Identifying more precisely which SES factors contribute to student performance. For example, Jencks and Phillips (1998) examined why the Black-White achievement gap persists even when obvious factors such as family income are controlled for and concluded that

family educational deficits persisted over several generations. Thus, the child's grandparents' education was an important predictor of educational achievement.

- Using records for individual students rather than school averages. Bingham, Heywood, and White (1991) reported on a multiple regression model to test the influence of 35 variables found in student records, including their individual characteristics and family background, previous schooling and achievement, and classroom attendance. These were used to develop an equation to predict achievement on the fifth-grade test. The residuals between the actual and predicted scores were then used to rank teachers.
- Measuring individual student gains rather than average test scores for a school. Most present ratings of school improvement compare a school's average score in one year with that school's score in the previous year. Any differences may simply reflect differences in the two student populations rather than a change in school performance. By contrast, annual testing using compatible tests allows schools to be rated by how well their students progress. The best-known system using changes in test scores alone to judge schools and teachers is the Tennessee Value Added Assessment System developed by Sanders and his associates. See Sanders and Horn (1998) for a description and Ross et al. (2001) for an example of its application to a school district.

Baker and Xu (1995) and Bock, Wolfe, Wolfe, and Fisher (1996) offer critiques.

- The development of more sophisticated statistical models, notably the hierarchical linear model (HLM) to take account of the interrelations between the levels of education: student, classroom, and school. In these, the equation used to predict student achievement includes a term calculated from the equation used to predict classroom results. The classroom prediction equation, in turn, includes a term for the school. Phillips and Adcock (1997) described the HLM model, and Meyer (1997) applied it to the case of annual testing.

Every model, including the one described in this article, has limitations. To keep these limitations in perspective, it may be useful to compare a theoretical model to common school ratings systems currently in use.

In many states, schools are rated by student performance on one or two tests. In Wisconsin, for instance, elementary schools are commonly judged by the percentage of their students receiving

proficient scores or better on either the third-grade reading test or on the fourth-grade reading and mathematics tests.

This approach creates an incentive for schools to concentrate their resources and best teachers on the class taking the important test. Because the shift of two or three students from one proficiency level to another may make a major impact on a school's rating, it also creates incentives to concentrate on the few students who are close to the line dividing proficiency levels.

Models that incorporate more of the available student performance data can make such gaming less attractive. By also incorporating SES data, such models make it less likely that poorly run schools in prosperous communities can gain high ratings on the strength of their students' relative prosperity.

Why haven't these models gained wider acceptance? They are commonly opposed by two groups: those opposed to standardized testing and those who fear that incorporating SES data justifies lower achievement by poor and minority students. But the lack of wider acceptance may also reflect the practical difficulties of applying many of these models. If the ideal model requires the collection of additional data, making hugely complex calculations using specialized software, or obtaining and matching individual student records, it may never be implemented.

In this article, I describe a statistical model that uses only publicly available data. The calculations required can be performed on a standard spreadsheet. The model can identify high-performing schools using all available test results as well as demographic data. I further describe its application to elementary schools in the Milwaukee Public Schools (MPS).

METHOD

My model consists of five steps:

1. Calculating regression coefficients that relate each school's average test scores to the poverty level of that school for each test given in a year.

2. Using these coefficients to calculate predicted test scores for each school based on the school's SES data.
3. Calculating the residuals—the differences between the actual and predicted test scores—for each school.
4. Converting these residuals into standardized residuals or effect sizes by dividing them by their standard deviation.
5. Averaging the effect sizes for each school over all the tests given in a year to get a school rating.

I applied this model to all elementary schools in the Milwaukee Public Schools (MPS). There are several reasons for choosing elementary schools over middle or high schools. First, the large number of schools, at least 112 depending on the year, minimizes the influence of any one school. Second is the increasing evidence of the importance of elementary education to later success.

A final reason for focusing on elementary schools centers on limitations stemming from the use of published data. MPS publishes average test score data for each school. It breaks out those averages by ethnic group if at least 10 students took the test. The data show a snapshot of average student performance at a particular time. Several middle and high schools have admissions requirements. Without data on incoming students, it becomes hard to separate the effect of the current school from that of earlier schools.[1]

Traditionally, the main use of regression analysis in education has been to explain student achievement by determining how various factors contribute to it. Yet demographic-based models do a strikingly poor job of explaining the differences in performance among schools, particularly in urban districts where many schools have high poverty rates. As discussed later, I found that demographic data typically accounted for only about 20% or 30% of the variation among schools.

It is this remaining 70-80% of student performance variation not explained by socioeconomic status that I wished to examine. It could have several possible explanations: (a) random variation, (b) some difference in student population not measured by the available demographic data, or (c) differences among schools such as instructional models or quality of school leadership.

Thus, my focus is less in measuring the influence of outside factors on achievement than in exploring and identifying differences

between schools that cannot be explained by these outside factors. By quantifying the influence of factors such as poverty on student achievement, my model aims to neutralize the effects of these factors and to isolate the school effects.

INDEPENDENT VARIABLES

Milwaukee Public Schools collects information on the socioeconomic characteristics of its students, including poverty, mobility, stability, and ethnic classification. In addition, school characteristics such as school size and the fraction of students taking a test are available.

Poverty is measured by the percentage of students participating in the free or reduced-priced lunch program. Enrollment in the subsidized meals program can depend on factors other than family income, including the social acceptability within the school of receiving free lunch and whether students like the food. Some schools may have true poverty rates somewhat higher or lower than the measures used in this study.[2] Despite its limitations, school lunch participation is the only measure of poverty readily available.

At the individual level, a student's lunch eligibility status normally takes one of only three values: free, reduced, or no participation. There is no indication of how much family income is above or below the eligibility limit. Theoretically, two schools could have identical participation rates and very different distributions of family income. In practice, however, with average school data, these differences probably wash out.

As found in many other studies, student performance is negatively correlated with poverty. As I describe later, in 2000-2001 MPS tested students in every grade from 2nd through 10th grade using the Terra Nova test. That test's scale scores are comparable from one test to the next. Figure 1 shows the scores for each year regressed against schools' poverty levels. The slope of each year's regression line is approximately 0.5, meaning that for every 1 percent increase in subsidized meal participation, the average scale score decreases by half a point. As a result, the average *third-grade* scale score of schools in which all students qualify for subsidized

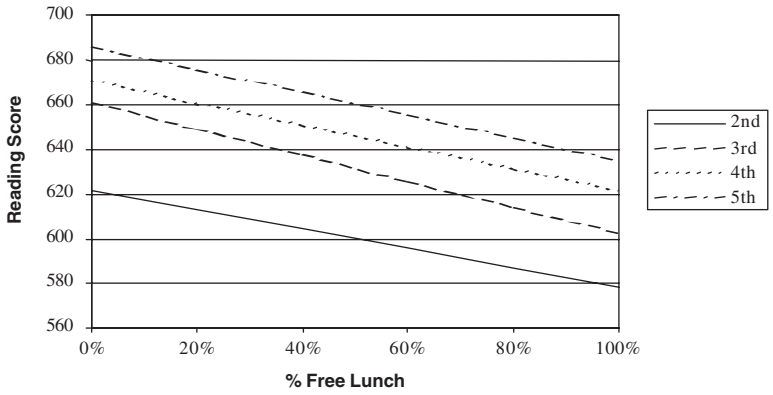

Figure 1 Reading Versus Free Lunch

meals equals the average *second-grade* score for schools where only half the children qualify.[3]

Another approach commonly taken to measuring the impact of poverty on student performance is through the coefficient of correlation *R*2. For most test scores, the *R*2 with poverty was between 0.2 and 0.4, implying that 20-40% of the variation in average scores could be explained by the variation in average school poverty rate.

As White, Reynolds, Thomas, and Gitzlaff (1993) documented, studies have found a wide variety of *R*2 values when looking at the connection between SES and achievement. Some of these differences may reflect variations in the distribution of poverty in the schools studied. For example, few MPS schools have poverty rates under 50%. A school system whose schools were distributed evenly throughout the poverty spectrum would have a higher value of *R*2 even if the underlying relationship of poverty to achievement were identical.[4]

Data are available on the *ethnic makeup* of each school. In addition, both average test scores and some demographic data (such as poverty and mobility) are available by ethnic group. I used the model to predict performance separately for African American and White students. A data limitation is the requirement that at least 10 students must be in the group to protect individual student privacy.

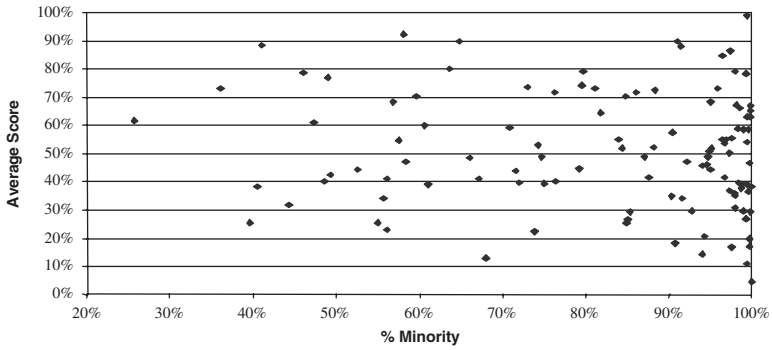

Figure 2 Scores Versus Minority Enrollment, 2000-2001

Some advocates believe that Black students learn better at schools with many White students. Conversely, others believe that Black students are better off with other Black students in an environment especially focused on their needs. My analysis supports neither position. For example, I found negligible correlation between 1996 and 1997 Black student test scores and either the percentage of White students in a school or that of Black students.

In Milwaukee Public Schools, minority enrollment is highly correlated with poverty ($r = 0.85$).[5] The resulting multicollinearity makes it difficult to separate the effect of minority enrollment on performance by using multiple regression. For most tests, the multiple regression coefficients relating percent minority to test scores were not statistically significant.

Rather than using multiple regression with poverty, minority enrollment, and other factors, I concluded that a cleaner approach is to first build the model using poverty as the only independent variable. The resulting ratings can then be regressed against minority enrollment, giving a measure of the additional effect of school ethnicity. Figure 2 shows a scatter plot comparing the ratings of MPS elementary schools based on poverty to their percentage of minority students. Correlation between minority enrollment and ratings is weakly negative ($r = –0.13$) but not statistically significant ($p = 0.19$).

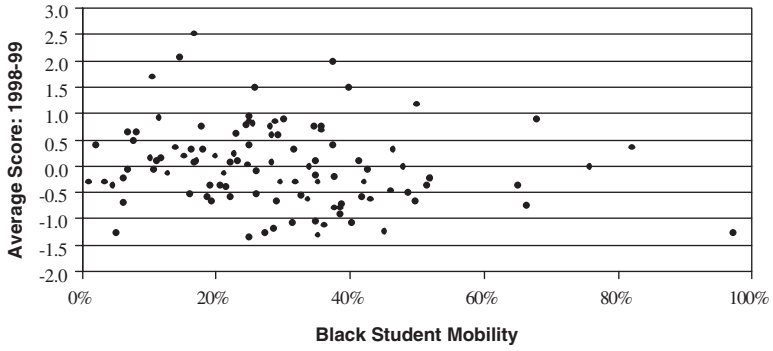

Figure 3 Scores Versus Mobility: African American Students

I also superimposed the results for schools based on their White students' performance and poverty level on top of a similar plot for Black students. The two sets of data seem to fall on the same continuum of performance.

In light of the considerable body of research documenting the Black-White achievement gap extending beyond the difference in poverty (see Jencks & Phillips, 1998, for example), I expected a stronger relationship. There are several possible explanations for the weak minority enrollment effect. One is that, in Milwaukee at least, the Black-White gap has largely disappeared and is simply driven by the poverty difference. Perhaps the difference would be greater at the middle and high school levels when factors such as peer pressure become greater. Another possible reason is that the small number of White students (around 16% of total enrollment) and their lower average poverty levels mask any Black-White gap.

Mobility is the percentage of students leaving during the school year.[6] *Stability* is the percentage of students who continue from one year to the next.[7] A stability rate of 100% means that all students return the following year; a rate of 0% means none return. The difference between the actual stability rate and 100% correlates closely with mobility ($r = 0.75$). Not unexpectedly, poverty also correlates with high mobility ($r = 0.49$).

Figure 3 compares average poverty-adjusted school scores for black students in 1998-1999 to the school mobility rates of Black students. Although, on average, scores drop as mobility rises, the

relationship is weak, with a $R2$ of only 0.04. The relationship is, however, statistically significant at the 95% level ($p = 0.04$).

A series of studies of mobility at MPS also found minimal impact on student performance. Thomas and White (1993) analyzed the effect of mobility on student performance, concluding that "the magnitude of the differences are 'vanishingly' small" (p. 1). In a study of busing at Milwaukee's P-5 schools, White and Thomas (1991) found that "busing has little or no meaningful relation to . . . student achievement" (p. 10). Heywood, Thomas, and White (1997) found "no evidence that mobility of classmates lowers achievement of stable students" (p. 354).

As with ethnicity, I expected a stronger relationship between mobility and student achievement. Even if mobility itself has no impact on achievement, high mobility can be a symptom of family economic or emotional stresses likely to depress achievement.

School performance on the state test is negatively correlated with *participation rates*, the percentage of students taking the test ($r = –0.2$). Most of the schools with test participation rates below 90% serve large numbers of Hispanic students who are exempted if their English language skills are judged insufficient. To the extent that difficulties in learning English are reflected in other academic limitations, exempting those students may give their schools a ratings advantage.

Spanish-speaking students exempted from the Terra Nova because of language are given the *Supera* test. Scale scores on the Supera are designed to be equivalent to those on the Terra Nova. Average Supera scores were lower than average Terra Nova scores at the same school, with the difference varying from 4 on reading to 18 on mathematics. Overall, it appears that exempting students because of language does give a boost to those schools' ratings.

This discussion hardly exhausts the list of independent variable candidates. Bingham et al. (1991), for example, identified more than 500 independent variables that could potentially relate to student performance.

Are more variables better? Multiple regression analysis allows the simultaneous use of numerous independent variables. I experimented with adding and subtracting independent variables to observe their effect on the model's predictive ability. Adding more

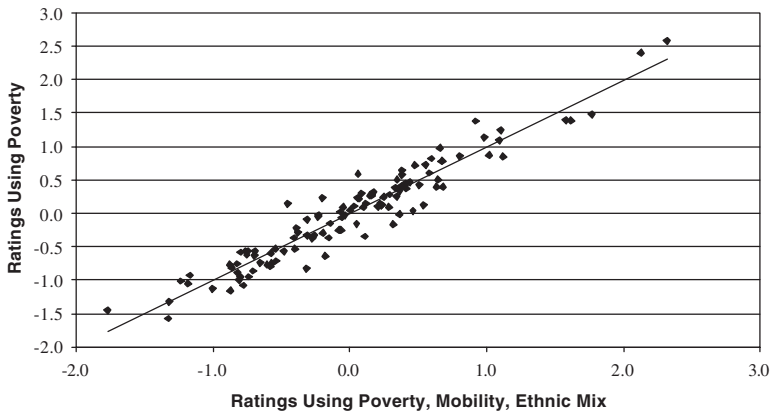

Figure 4 Comparison of School Ratings, 1999-2000

independent variables sometimes led to improvements in such measures of fit as the adjusted *R*2, but the improvements were slight. It also often led to a result in which none of the independent variables was statistically significant, reflecting their mutual correlations.

In deciding which independent variables to include, I examined whether their inclusion would significantly change the schools' ratings. Figure 4 shows 1999-2000 school ratings based on two different sets of independent variables. Each dot represents a school. The horizontal scale shows ratings using poverty, mobility, percent Black, percent White, and percent minority students as independent variables. The vertical axis shows ratings using only the poverty rate. Increasing the number of independent variables results in little change in the rankings of the schools, particularly for schools at the extremes. The basic conclusion remains: Some schools do very much better than others under any reasonable version of the model.

This robustness of results stems in part from multicollinearity of the demographic data. On average, schools with higher poverty have more minority students and suffer higher rates of mobility over the school year and lower rates of stability from one year to the next. To a significant extent, the demographic factors are interchangeable.

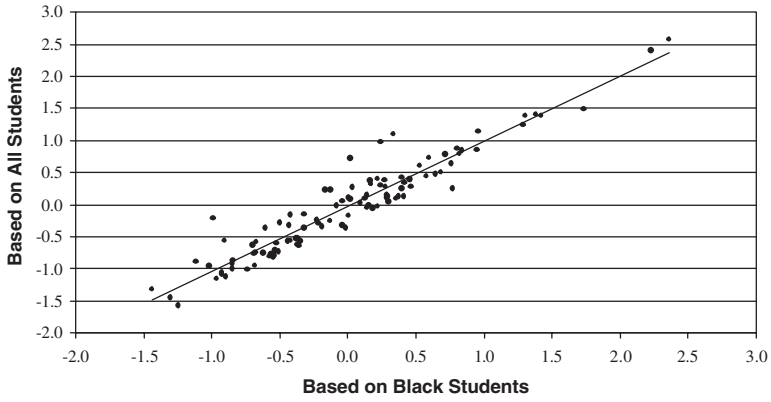

Figure 5 Ratings Based on All Versus Black Students, 1999-2000

I also compared school ratings based on all students in the school to models based on single ethnic groups. I developed ratings based on each ethnic subgroup's average school test scores and poverty rate. Because test scores are not supplied if fewer than 10 students in the group took the test, this approach limited the number of schools included, particularly with Hispanic and White students. If schools were missing data on only a few tests, I substituted average data.

Figure 5 compares schools' ratings based on all students to the same schools' ratings based only on their African American students. In most cases, the school ratings were similar.

As a result, most ratings discussed in this article use schools' overall poverty as the independent variable, measured by the percentage of students who qualify for the subsidized free or reduced-priced lunch program. Except when noted, the ratings are based on the performance of all students rather than of particular ethnic groups.

DEPENDENT VARIABLES

I applied the regression model to each of the tests given at all elementary schools. For most of the period covered in this study, Milwaukee Public Schools gave a mixture of state tests and district-

Table 1
Former MPS Assessment System

Grade	*State/District*	*Subjects*
Third	Wisconsin	Reading
Fourth	Wisconsin	Reading, language arts, mathematics, science, social studies
	MPS	Writing performance
Fifth	MPS	Writing and science performance

designed performance assessments, as shown in Table 1. In earlier years, MPS also gave the fifth-grade Iowa Test of Basic Skills in mathematics.

The Wisconsin 4th-, 8th-, and 10th-grade Knowledge and Concepts Exams use the Terra Nova tests from CTB/McGraw-Hill. The Wisconsin third-grade reading comprehension test was developed by the State Department of Public Instruction and is not calibrated to other Wisconsin tests. All school districts in the state must give these tests. The state commonly reports the results as the percentage of students scoring in one of four proficiency categories: minimal, basic, proficient, or advanced.

In addition to the state-mandated tests, MPS gave locally developed performance assessments in writing, science, and mathematics. In these assessments, students scoring a certain number of points were judged proficient.

During the 2000-2001 school year, MPS implemented a new testing plan centered on the Terra Nova test; it is summarized in Table 2. Because Wisconsin also uses the Terra Nova, annual gains made each year for each student can be calculated. Starting with 2000-2002, MPS dropped all second-grade tests and all third-grade Terra Nova tests except in mathematics.

The MPS and Wisconsin Terra Nova exams report student results in three different formats: criterion-referenced scores (minimal, basic, proficient, or advanced), nationally normed percentiles, and scale scores. Using item-response theory, scale scores are designed to be consistent throughout levels of the Terra Nova so that a score of 560 on the first-grade reading exam is equivalent to a 560 on the third-grade examination. Giving the exam every year will allow comparisons of student progress over time.

Table 2
New Testing Plan Starting 2000-2001

Grade	*State/District*	*Milwaukee Public Schools*
Second Grade	MPS	Terra Nova: reading, language arts, math (dropped in subsequent years)
Third Grade	Wisconsin	Reading
	MPS	Terra Nova: reading, language arts, math (only math in subsequent years)
Fourth Grade	Wisconsin	Reading, language arts, mathematics, science, social studies
	MPS	Writing performance
Fifth Grade	MPS	Terra Nova: reading, language arts, math Writing performance

For each test, I compare the result predicted for each school based on its poverty rate to the school's actual score. I then calculate the residuals, which are the gaps between actual and predicted test results. I convert these to effect sizes (also called standardized residuals, or *z* values) by dividing each value by the standard deviation. A school achieving at exactly its predicted level has an effect size of 0. Schools with actual scores exceeding their predicted level have positive effect sizes, and those whose actual scores fall below prediction have negative values. I then average the effect sizes over all tests to get overall school ratings for the year.

This model allows the use of any type of test scores, scale scores, proficiency percentages, or national percentiles as long as they are numerical. Scores based on proficiencies may, however, be the least stable because the movement of a small number of students across the dividing line between proficiencies can have a large effect on a school's ratings.

To make the ratings easier to grasp, I also convert them to letter grades. I chose grade cutoff points so that on individual tests, 20% of the schools would fall into each grade between A and E.

Statistical theory predicts that when values are averaged over a number of tests to get an overall school score, average scores will concentrate at the middle. Yet, as shown in Figure 6, the concentration is much smaller than predicted by theory. This phenomenon

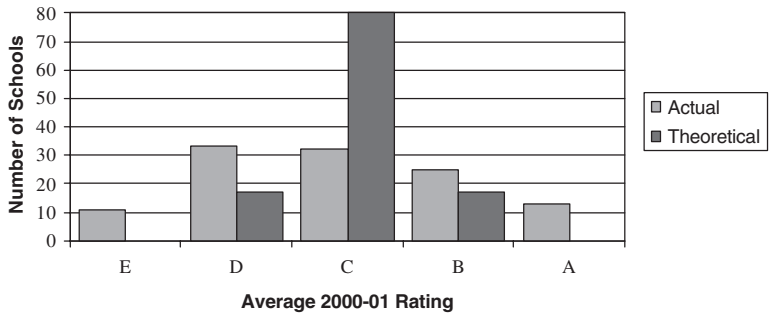

Figure 6 Schools in Categories

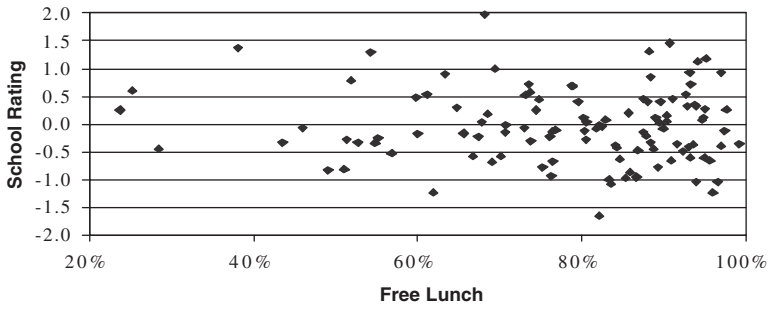

Figure 7 2000-2001 School Ratings Versus Poverty Rate

reflects the strong relationship between a school's results on one test and its results on others.

Figure 7 shows the results for all MPS elementary schools in all tests given in the 2000-2001 school year. Each dot represents an MPS school. The vertical scale in this figure shows the overall score for each school for that year. The horizontal scale represents the average school poverty rate as represented by the percentage of students qualifying for subsidized meals. Most schools are on the right-hand side of this graph, reflecting high poverty among MPS school children.

The model is very flexible. It can be readily modified to apply to any groups of schools as long as accurate demographic data, partic-

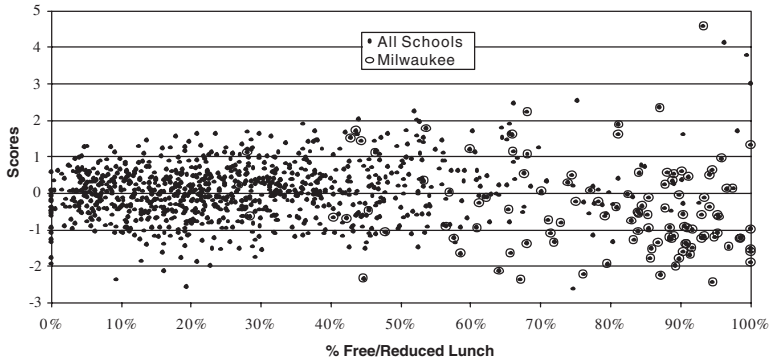

Figure 8 2001-2002 Wisconsin Elementary School Scores

ularly on poverty and test scores, are available. For example, Figure 8 shows the results of applying the model to all elementary schools in Wisconsin. Each dot represents a Wisconsin public school. The dots with circles are schools in the Milwaukee Public Schools. The horizontal scale represents the poverty rate in the school as reported by the Wisconsin Department of Public Instruction (DPI). Not surprisingly, most MPS schools are on the right side of this graph, reflecting the high poverty rate typical of MPS schools. (The high ratings of the non-MPS schools in the upper-right-hand corner seem to reflect errors in their reported poverty levels.)

RESULTS

STABILITY OVER TIME

If school ratings reflected random variation, one would expect that there would be little stability from one year to the next. Conversely, to the extent that they represent some fundamental school characteristic, they should change only as the school changes.

Several approaches can be taken toward measuring time stability. One is to examine how one year's top schools did in other years. Table 3 shows the top 10 performing schools for all students based

Table 3
Top 2000-2001 Schools Over Time

	1996-1997	*1997-1998*	*1998-1999*	*1999-2000*	*2000-2001*	*2001-2002*
School A	A	A	A	A	A	A
School B	C	D	B	A	A	A
School C	C	D	B	B	A	A
School D	A	A	B	B	A	A
School E	C	A	A	A	A	A
School F	B	C	A	B	A	A

Table 4
Bottom 2000-01 Schools Over Time

	1996-1997	*1997-1998*	*1998-1999*	*1999-2000*	*2000-2001*	*2001-2002*
School K	D	D	D	D	E	E
School L	E	D	B	A	E	D
School M	E	D	E	E	E	E
School N	C	D	D	D	E	E
School O	E	D	E	C	E	C
School P	E	C	D	E	E	D

on 2000-2001 scores. Four years earlier, the majority was rated either A or B. On the whole, schools doing well in one year do well in other years. This result supports the hypothesis that these schools have learned to significantly affect student performance.

The bottom 10 performers in 2000-2001 are shown in Table 4. For most of these schools, their problems could be seen years earlier.

A second approach uses a scatter plot to compare schools' performance in two different years. In Figure 9, I compare schools' rankings in 1996-1997 to those in 2000-2001. Each point represents an MPS elementary school. The horizontal scale shows schools' 1996-1997 rankings and the vertical scale shows their 2000-2001 rankings. Some schools moved up and others down, but there was an overall relationship between a school's ratings in the two years. The relationship is statistically significant (p 0.000001) with an $R2$ of 0.18. It may be worthwhile noting that there is no overlap between the two cohorts of students tested, because the first

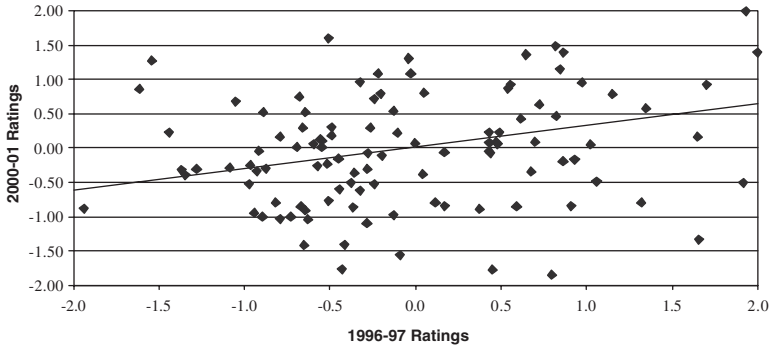

Figure 9 1996-1997 Ratings Versus 2000-2001 Ratings

group would have graduated from elementary school by the time of the second tests.

A third approach is to apply statistical measures to examine whether the results could have resulted from random variation. For example, consider two groups of MPS elementary schools. The first consists of the 40 schools scoring highest in the 1996-1997 testing season. In the second are that year's bottom 40 schools. The first group, on average, performed significantly better than the second on the 2000-2001 tests. Analysis of variance (ANOVA) generates a *p* value of 0.0000425.

In spring 2001, MPS tested students across grades using the Terra Nova test. Figure 10 shows the average reading scores for second through fifth grade. I selected four groups of schools based on their 1996-1997 ratings: schools rated A, those rated A or B, those rated D or E, and those rated E. The average*third-grade* student in a school placed in the top group based on its ratings five years earlier read at about the same level as the average *fifth-grade* student in a school rated at the very bottom five years before.

The results for mathematics are similar, as shown in Figure 11. Again, grouping schools by their scores from five years earlier leads to real differences in student achievement on tests given years later. The results for language arts (not shown) are similar to those for reading. The differences between groups cannot be attributed to a particularly able cohort of students, because students tested in

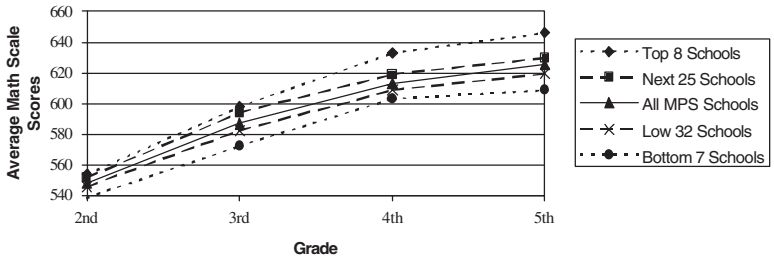

Figure 10 Reading Scores, 2000-2001

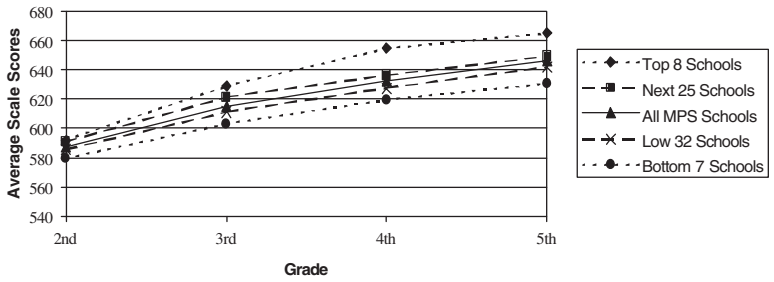

Figure 11 Math Scores, 2000-2001

1996-1997 would have been promoted out of elementary grades long before the 2001 Terra Nova elementary tests.

In sum, although a few schools make substantial movements up or down in the ratings, overall ratings remain quite stable.

My model identifies differences in the ratings of schools serving apparently similar populations. These differences remain after all available data on school populations are used to predict school performance. Two other factors besides differences in the school's programs and practices could contribute to the school ratings: (a) unmeasured differences in school populations and (b) gaming the results.

UNMEASURED DIFFERENCES IN STUDENT POPULATION

If the measured differences I discuss above do not account for most of the differences among schools, could some other external

factor not reflected in measured data significantly affect school performance? One possibility is that schools whose student populations are similar socioeconomically might nevertheless attract students with different abilities or readiness to learn.

One way this could happen is through a kind of creaming process. Through explicit admissions requirements or more subtle policies, a school's entering student population may have advantages that do not show up in the demographic data. At the elementary level, only one school has an explicit admission policy. Among the others, citywide specialty schools might be regarded as having implicit admissions policies. Because they typically fill early, their students likely come from families who set a high value on education. The average socioeconomic status of students at these schools is indeed higher. Once I adjusted the results for poverty, however, the specialty school performance advantage disappears. The average 2000-2001 rating for citywide schools is virtually identical to the average for all schools.

Another possibility is that there is something unique about the population a school serves, some cultural difference that does not show up in demographic data. Certainly, this is possible. One can point to several historical examples of poverty-stricken immigrants who set such a high priority on education that their children would make any school look good. There is no anecdotal evidence of that happening in present-day Milwaukee. In fact, several schools in the top 10 are adjacent to schools in the bottom 10, often drawing children from the same neighborhoods.

GAMING THE TESTS AND DATA INTEGRITY

Finally, the possibility remains that some schools' scores are inflated due to strategies ranging from excessive "teaching to the test" to outright cheating. There are many stories of how schools try to improve their scores. Some of these strategies are clearly desirable, such as identifying students falling behind in reading or mathematics and finding ways to meet their needs. Others, such as encouraging low-performing students to skip test day, undermine the accuracy of the results. As tests take on more consequences, issues of questionable testing practices will gain more urgency.

It is likely that the rules for giving tests will become both stricter and more consistent. Compared to the ACT or SAT college admissions tests, the rules on state testing are quite flexible. For example, schools are given a 1-month window to administer tests, allowing teachers to view the tests and perhaps tailor their classes to the material on the tests.

Part of the solution may come out of increased testing. Inconsistencies between tests given over time or different tests given at the same time may help raise warning flags. For example, if individual student scores in a class or school subsequently decline, that may be a sign that the previous scores were inflated. Unusually high or low dispersions in test results may be an indication of manipulation. The possibility that there will be deliberate distortion of test results is another reason to incorporate all available test results into the model.

The independent variables—the measures of poverty—may also have errors. In Figure 8 showing Wisconsin school results, for example, two non-MPS schools are rated very high. The state's data show these two schools having poverty rates near 100%. Yet the previous year's data show their poverty around 10%.

Before decisions about a school are made based on any rating system, the accuracy of the data used to rate them should be double-checked.

PREDICTING FUTURE STUDENT PERFORMANCE

Perhaps the most important test of any school is how well its students perform after they leave the school. Do the students from the higher rated schools continue to perform above expectations when they move on to sixth grade? If the students' achievement falls to the level predicted by their socioeconomic status, that would imply either that whatever the successful schools do does not have lasting value or that their previous results were inflated.

As mentioned earlier, Milwaukee Public Schools recently started testing students annually using the Terra Nova reading, language arts, and mathematics tests. Using a relational database of student records, I calculated average sixth-grade test scores for each elementary school based on the school the students attended

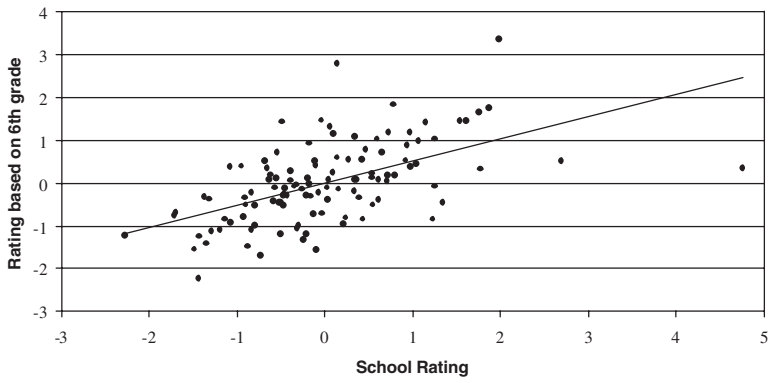

Figure 12 One Year After

in fifth grade. I then calculated the predicted scores for each school based on its free lunch and used that to calculate a sixth-grade rating for each school.[8]

Figure 12 compares the ratings of the elementary schools based on the sixth-grade scores their students achieved in the spring of 2002 to the ratings of the schools that were calculated using all tests given at the school in the 2000-2001 school year. There is a positive relationship between performance in the elementary school and how the students did when they entered middle school. The slope of the trend line is 0.52.

The school shown as a point on the far right of Figure 12 is an example of a school whose test results deserve more checking. On average, the average fourth- and fifth-grade scores at this school are very high, but students' scores fall to an average level once they move to another school.

A COMPARISON OF SCHOOL RATING SYSTEMS

Although opposition to rating schools is likely to continue, the use of ratings will probably increase, inspired both by the desire to identify schools that "beat the odds" and the requirement under the federal No Child Left Behind Act (NCLBA) to identify schools needing intervention. In this section, I compare school ratings generated by my model to two such rating systems.

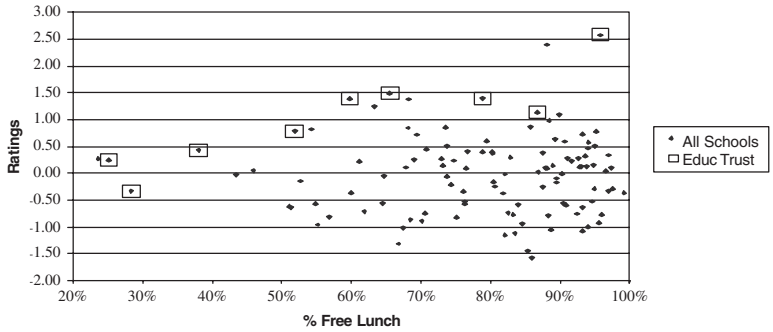

Figure 13 Ratings for 1999-2000

The Education Trust (2001) has published a series of reports under the title *Dispelling the Myth*. Accompanying these reports are lists of schools that have high test scores despite high poverty or high minority enrollment. Jerald (2001) included such a list based on reading or mathematics tests in 1999-2000. Figure 13 shows the ratings of these schools given by my model. It is evident that the lower the poverty rate in a school, the more likely it is to be included in the list.

The two schools with the highest poverty that made the list underscore the challenges of achieving consistently high scores when most of the students suffer from poverty. As already noted, students from one of these schools suffer a marked decline in test scores when they switch schools. The other school was included in the list because of high fourth-grade math and reading scores in 1999-2000. It other test scores that year were, however, at the low end, as were all its test scores in other years.

In the second case, Wisconsin released its list of *schools needing improvement* under the NCLBA based on 2000-2001 testing (see Figure 14). These are commonly dubbed *failing schools*. Wisconsin used the fourth-grade reading and math tests to arrive at this designation, as well as changes in the scores from one year to another.[9] In this case, the chance of inclusion increases with poverty. The relationship between poverty and the need for improvement is particularly evident when looking at all Wisconsin schools.

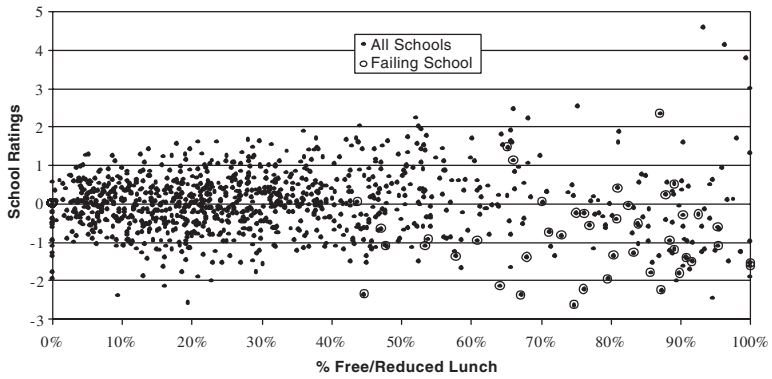

Figure 14 2001-2002 Wisconsin Elementary School Ratings

Several schools rated quite highly in my model make the failing school list. Their inclusion reflects both the limited number of tests used by the state and its use of comparisons between average scores in one year and the next. In some cases, schools on the list had substantially higher test scores than schools not on the list.

CONCLUSIONS

A model such as the one I describe has both uses and limitations. The importance of the limitations depends in many cases on how the model is used. Here are five possible uses:

1. Identifying high-performing schools for examination and replication. One group of schools consistently outperforms the levels predicted by their poverty rates. Another group of schools consistently fails to achieve the predicted levels. Systematic school improvement could result from discovering the successful schools' secrets in encouraging student achievement and finding ways to transfer those secrets to other schools.

This bottom-up approach to school reform contrasts with the more common top-down approach in which *whole-school reform models* developed at universities or think tanks are applied to

schools. The attempt to discover and learn from high-performing schools is not new. Many of the models aimed at identifying such schools depend, however, on the results of one or two tests. This approach can lead to instability so that a school identified as high-performing one year may be listed for intervention a year or so later. By contrast, my model gives a far more stable list of high-performing schools.

2. Developing school report cards. Increasingly, report cards are used to rate schools. Often, they are accompanied by rewards and punishments. Recent amendments to federal education law accelerate that trend. As the stakes in these report cards rise, equity issues will increase, particularly if every school slated for intervention serves low-income students.

Whether schools are rated by raw scores on one or two tests or by a model such as that I describe here will become of increasing importance. If the former, it is likely that most schools rated low will serve low-income children and that few of the schools with high ratings will have low-income or minority populations.

Whether adjustment should be made for a school's minority population is often politically sensitive. Experience with my model suggests that the penalty to high-minority schools of ignoring ethnicity is moderate as long as poverty is incorporated into the model.

3. Measuring the impact of outside factors, such as poverty, race, or mobility, that affect school performance so that society can address them both inside and outside schools. This has been the thrust of much previous work on SES and school performance. Unfortunately, it can also result in shifting the focus away from what schools can do to raise student achievement.
4. Evaluate the impact of programs introduced in a group of schools. Do the affected schools increase their ratings compared to others? The use of the model depends on participation by a sufficient number of schools for statistical significance but not so many as to lose a control group. In a forthcoming publication (Thompson, 2003), I report on the use of this model to evaluate the effects of three programs introduced during the span of this study. For systemwide reforms, one would need to look for change in the absolute results of tests, such as the NAEP or statewide tests. Unfortunately, it is common for states to change their tests, making year-to-year comparison difficult.

5. Gain insight on policy issues. I developed an early version of this model because as a member of the Milwaukee school board, I grew frustrated at the lack of information on the impact of policies.

In several instances, the model helped in making decisions:

- MPS has moved to give schools more authority over budget and hiring decisions. One group of support specialists lobbied the board intensely for a policy that would require every school to hire one of them. Regression analysis, however, showed no relationship between school ratings and whether such a specialist was at the school. The board did not pass the proposal.
- As the number of Black students increased, most of the principals appointed to predominantly Black schools were African American, reflecting a belief that a role model would help student achievement. Yet my model showed that schools with Black principals had substantially lower Black student performance than those with White principals. In its rush to appoint more Black principals, it appeared that MPS neglected issues of selection, support, and evaluation. The Black principals on average were younger, were less experienced, and lacked the informal support network that more senior principals enjoyed. Recently, MPS extensively revamped its principal selection process and increased the training and support given to new principals.
- Some advocates of expanded busing argued for a policy to bus all students who move back to their school, a very expensive option. The weak relationship between mobility and achievement suggests that money spent on such transportation might be better directed toward educational programs.

This model can be applied to any group of schools as long as sufficient performance and demographic data are available. Additional data (such as gain scores) can be readily incorporated into the model as they become available.

NOTES

1. In spring 2001, MPS introduced a new annual testing program that will permit the calculation of value-added scores. These data can be used to calculate the average incoming scores of students who enter middle and high schools. Such an analysis could not, of course, use only published data. At the elementary level, the exclusive use of gain scores leaves early classes unmeasured. Under the current testing plan, for example, the first gains could be calculated in MPS for the fifth grade in reading and in the fourth grade in the other two subjects.

2. For example, the high school participation rate is substantially lower than the elementary school rate.

3. At the upper levels, the slope is even steeper: around 0.7 for middle school and 1 for high schools. Some of this increase may reflect admissions standards.

4. I simulated this relationship by using the regression coefficient for fourth-grade reading scores for 2000-2001, superimposing random variation using the standard deviation of the residuals. For a simulated school system with average school free-lunch participation varying uniformly from 0% to 100%, the value of $R2$ was 0.52. For one with the free-lunch distribution identical to that at MPS schools, $R2$ dropped to 0.29.

5. Unless otherwise specified, all statistics are based on data for all students in 2000-2001.

6. The *percent mobility* is officially defined as the sum of students who enter and leave a school between the third Friday of September and the last day of school in June, divided by the official Third Friday September enrollment.

7. The 1-year stability rate for students in grades other than the highest is measured from the Third Friday September enrollment to the next Third Friday September count.

8. Because this calculation depended on using individual student records, it could not be done using published data. I did not, however, incorporate these results into the model. Rather, they were used to help confirm that the model's rating reflect real differences in schools.

9. The rather complex calculations are shown on the department's Web site at http://www.dpi.state.wi.us/oea/kce998.html (Benson, 1998).

REFERENCES

Baker, A. P., & Xu, D. (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.

Benson, J. (1998). Re: Statewide accountability policy update. Retrieved December 14, 2003, from http://www.dpi.state.wi.us/oea/kce998.html

Bingham, R. D., Heywood, J. S., & White, S. B. (1991). Evaluating schools and teachers based on student performance: Testing an alternative methodology. *Evaluation Review 15*(2), 191-218.

Bock, R. D., Wolfe, R. D., Wolfe, R., & Fisher, T. H. (1996). *A review and analysis of the value added assessment system* . Nashville: Tennessee Controller of the Treasury.

Borsuk, A. J. (2002, August 7). 38% of MPS schools face sanctions: State lists reading, math failures in 64 Milwaukee schools, 20 others. *Milwaukee Journal Sentinel.*

Clotfelter, C., & Ladd, H. (1996). Recognizing and rewarding success in public schools. In H. Ladd (Ed.), *Holding schools accountable* (pp. 23-64). Washington, DC: Brookings Institution.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Department of Health, Education and Welfare.

Education Trust. (2001). *Dispelling the myth*. Retrieved December 14, 2003, from http://www2.edtrust.org/edtrust/dtm/

Fitz-Gibbon, C. T., & Tymms, P. (2002). Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives*, *10*(6). Retrieved January 30, 2002, from http://epaa.asu.edu/epaa/v10n6/

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

Heywood, J. S., Thomas, M., & White, S. B. (1997). Does classroom mobility hurt stable students? An examination of achievement in urban schools. *Urban Education*, *32*, 354-372.

Innes, T. C., & Cormier, W. H. (1973, March). *The prediction of achievement means of schools from non-school factors through criterion scaling* Paper presented at the annual meeting of American Educational Research Association, New Orleans.

Jencks, C., & Phillips, M. (1998). *The black-white test score gap* Washington, DC: Brookings Institution.

Jerald, C. D. (2001). *Dispelling the myth revisited: Preliminary findings from a nationwide analysis of high-performing schools*. Washington, DC: Education Trust.

Meyer, R. H. (1997). Value-added indicators of school performance: a primer, *Economics of Education Review*, *16*(3), 283-301.

Phillips, G. W., & Adcock, E. P. (1997). Measuring school effects with hierarchical linear modeling: data handling and modeling issues. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Ross, S. M., Sanders, W. L., Wright, S. P., Stringfield, S., Wang, L. W., & Alberg, M. (2001). Two- and three-year achievement results from the Memphis restructuring initiative. *School Effectiveness and School Achievement*, *12*(3), 323-346.

Sanders, W. L., & 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.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Thomas, M., & White, S. B. (1993). Do students who remain at the same school do better? Unpublished manuscript, University of Wisconsin–Milwaukee Urban Research Center.

Thompson, B. R. (2003). *Evaluating three programs using a school effectiveness model: Direct instruction, target teach, and class size reduction*. Manuscript submitted for publication.

Thrupp, M. (2001). Sociological and political concerns about school effectiveness research: Time for a new research agenda. *School Effectiveness and School Improvement*, *12*(1), 7-40.

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.

White, S. B., & Thomas, M. (1991). Busing for stability and student achievement in P-5 program Milwaukee schools. Unpublished manuscript, University of Wisconsin–Milwaukee Urban Research Center.

Bruce R. Thompson is professor and director of the Master of Science in Engineering Management program at the Rader School of Business, Milwaukee School of Engineering. A former president of the Milwaukee Board of School Directors, his research interests have been in using statistical and other mathematical models to better understand the effectiveness of institutions and programs.