Are "Failing" Schools Really Failing? Using Seasonal Comparison to Evaluate School Effectiveness

Douglas B. Downey, Paul T. von Hippel and Melanie Hughes

Sociology of Education 2008 81: 242
DOI: 10.1177/003804070808100302

Are "Failing" Schools Really Failing? Using Seasonal Comparison to Evaluate School Effectiveness

Douglas B. Downey
Paul T. von Hippel
The Ohio State University
Melanie Hughes
University of Pittsburgh

To many, it seems obvious which schools are failing—schools whose students perform poorly on achievement tests. But since evaluating schools on achievement mixes the effects of school and nonschool influences, achievement-based evaluation likely underestimates the effectiveness of schools that serve disadvantaged populations. In this article, the authors discuss school-evaluation methods that more effectively separate school effects from nonschool effects. Specifically, the authors evaluate schools using 12-month (calendar-year) learning rates, 9-month (school-year) learning rates, and a provocative new measure, "impact"—which is the difference between the school-year learning rate and the summer learning rate. Using data from the Early Childhood Longitudinal Study of 1998–99, the authors show that learning- or impact-based evaluation methods substantially change conclusions about which schools are failing. In particular, among schools with failing (i.e., bottom-quintile) achievement levels, less than half are failing with respect to learning or impact. In addition, schools that serve disadvantaged students are much more likely to have low achievement levels than they are to have low levels of learning or impact. The implications of these findings are discussed in relation to market-based educational reform.

Market-based reforms pervade discussions of current educational policy in the United States. The potential for markets to promote efficiency, long recognized in the private sector, represents an attractive mechanism by which to improve the quality of public education, especially among urban schools serving poor students, where inefficiency is suspected (Chubb and Moe 1990; Walberg and Bast 2003). Both the rapid growth of charter schools (Renzulli and Roscigno 2005) and the emphasis on accountability in the No Child Left Behind (NCLB) Act are prompted by the belief that when parents have information about the quality of schools, accompanied by a choice about where to send their children, competitive pressure will encourage administrators and teachers to improve schools by working harder and smarter.

Critical to the success of a market system is the need for consumers (i.e., parents) to have good information about the quality of services (i.e., schools) because market efficiency is undermined if information is unavailable or inaccurate (Ladd 2002). Toward this end, the

NCLB requires states to make public their evaluations of schools, addressing the need for information on quality to be easily accessible.

But do states' usual evaluations provide valid information on school quality? Are the schools that are designated as "failing," under current criteria, really the least effective schools? Under most current evaluation systems, "failing" schools are defined as schools with low average achievement scores. The basis for this definition of school failure is the assumption that students' achievement is a direct measure of school quality. Yet we know that this assumption is wrong. As the Coleman report and other research highlighted decades ago, achievement scores have more to do with family influences than with the quality of schools (Coleman et al. 1966; Jencks et al. 1972). It follows that a valid system of school evaluation must separate school effects from nonschool effects on children's achievement and learning.

Since the 1966 Coleman report, sociologists' contributions to evaluations of schools have been less visible, with current educational legislation dominated by ideas from economics and, to a lesser extent, psychology. In this article, we show how ideas and methods from sociology can make important contributions in the effort to separate school effects from nonschool effects. Specifically, we consider evaluating schools using 12-month (calendar-year) learning rates; 9-month (school-year) learning rates; and a provocative new measure, "impact," which is the difference between the school-year learning rate and the summer learning rate. The impact measure is unique in that its theoretical and methodological roots are in sociology.

One may expect that the method of evaluation would have little effect on which schools appear to be ineffective. After all, schools that have been identified as failing under achievement-based methods do look like the worst schools. They not only have low test scores, but they tend to have a high turnover of teachers, low levels of resources, and poor morale (Thernstrom and Thernstrom 2003). Yet we will show that if we evaluate schools using learning or impact—that is, if we try to isolate the effect of school from nonschool factors on students' learning—our ideas about failing schools change in important ways. Among schools that are failing under an achievement-based criterion, less than half are failing under criteria that are based on learning or impact. In addition, roughly one-fifth of schools with satisfactory achievement scores turn up among the poorest performers with respect to learning or impact.

These patterns suggest that raw achievement levels cannot be considered an accurate measure of the effectiveness of schools; accurately gauging school performance requires new approaches. Achievement-based indicators of school effectiveness are subject to considerable error and have limited utility for helping schools to improve. Evaluating schools on learning or impact would provide better information to parents and lead to a more efficient educational market.

THREE MEASURES OF SCHOOL EFFECTIVNESS

In this section, we review the most widely used method for evaluating schools—achievement—and contrast it with less-often-used methods that are based on learning or gains. We discuss the practice of using students' characteristics to "adjust" achievement or gains and highlight the problems that are inherent in making such adjustments. We then introduce a new evaluation measure that we call *impact*—which measures the degree to which a school's students learn faster when they are in school (during the academic year) than when they are not (during summer vacation).

Achievement

Because success in the economy, and in life, typically requires a certain level of academic skill, the NCLB generally holds schools accountable for their students' levels of achievement or proficiency. At present, the federal government allows each state to define proficiency and set its own proficiency bar (Ryan 2004), but the NCLB provides

guidelines about how proficiency is to be measured. For example, the NCLB requires all states to test children in math and reading annually between Grades 3 and 8 and at least once between Grades 10 and 12. In addition, states must test students in science three times between Grades 3 and 12. As one example of how states have responded, the Ohio Department of Education complies with the NCLB by using an achievement-bar standard for Ohio schools that is based on 20 test scores spanning different grades and subjects, as well as two indicators (attendance and graduation rates) that are not based on test scores.

In some modest and temporary ways, the NCLB acknowledges that schools serve children from unequal nonschool environments and that these nonschool influences may have some effect on children's achievement levels. For example, schools with low test scores are not expected to clear the state's proficiency bar immediately; they can satisfy state requirements by making "adequate yearly progress" toward the desired level for the first several years. (The definition of "adequate yearly progress" varies by state.[1]) In this way, the legislation recognizes that schools that serve poor children will need some time to catch up and reach the proficiency standards that are expected of all schools. By 2013–14, however, all schools are expected to reach the standard. More important, for our purposes, the schools that "need improvement" are identified mainly on the basis of their achievement scores.

The main problem with evaluating schools this way is that achievement tests do not adequately separate school and nonschool effects on children's learning. It is likely that a schools' test scores are a function not just of school practices (e.g., good teaching and efficient administration), but of nonschool characteristics (e.g., involved parenting and high-resource neighborhoods). It is unclear, therefore, the extent to which schools with high test scores are necessarily "good" schools and schools with low test scores are necessarily failing. Students are not randomly assigned to schools, so there is considerable variation in the kinds of students who attend different schools. Thus, when one evaluates schools, the challenge is to measure the value that schools add *independent of the widely varying nonschool factors that also influence achievement.*

Sociologists have documented extensively the importance of the home environment to children's development, along with the substantial variation in children's home experiences. As one example of how much home environments vary in cognitive stimulation, Hart and Risley (1995) observed that among children aged 6 months to 3 years, those whose families were on welfare had 616 words per hour directed to them compared to 1,251 words for children of working-class parents and 2,153 words for children of professional parents. Given such varying exposure to language, it is not surprising that large gaps in skills can be observed among children at the beginning of kindergarten. For example, 18 percent of children who entered kindergarten in the United States in the fall of 1998 did not know that print reads from left to right, did not know where to go when a line of print ends, and did not know where the story ends in a book (West, Denton, and Germino-Hausken 2000). At the other end of the spectrum, a small percentage of kindergarten entrants could already read words in context (West et al. 2000).

Of course, widely varying skills among children would not be so problematic for the goal of measuring school effectiveness if children's initial achievement levels were randomly distributed across schools, but even at the beginning of kindergarten, achievement levels differ substantially from one school to another (Downey, von Hippel, and Broh 2004; Lee and Burkam 2002; Reardon 2003). At the start of kindergarten, 21 percent of the variation in reading test scores and 25 percent of the variation in math test scores lies between, rather than within, schools (Downey et al. 2004). In other words, substantial differences in school achievement levels are observable even before schools have a chance to matter. Obviously, these variations are not a consequence of differences in school quality, but represent the fact that schools serve different kinds of students.

Although children's achievement is clearly influenced by both school and nonschool fac-

tors, achievement-based methods of evaluating schools assume that only schools matter. As a result, the burden of improvement is disproportionately placed on schools that serve children from poor nonschool environments, even though it is not clear that these schools are less effective than are schools that serve children from advantaged environments. Although some schools that serve disadvantaged populations may actually be poor-quality schools, without separating school effects from nonschool effects, it is difficult to make this evaluation with confidence. These criticisms of achievement-based measures of school effectiveness are, by now, well established in the social science community (cf. Scheerens and Bosker 1997; Teddlie and Reynolds 1999).

Learning

One way to measure school effectiveness that begins to address differences in nonschool factors is to gauge how much students learn in a year, rather than where they end up on an achievement scale. The advantage of an approach based on learning is that schools are not rewarded or penalized for the achievement level of their students at the beginning of the year. Under a learning-based evaluation system, schools that serve children with initially high achievement would be challenged to raise students' performance even further, while schools that serve disadvantaged students could be deemed "effective" if the students made substantial progress from an initially low achievement level, even if their final achievement level was still somewhat low.

One example of a learning-based evaluation system is the Tennessee Value Added Assessment System (TVAAS), implemented by Tennessee in 1992 to assess its teachers and schools (Sanders 1998; Sanders and Horn 1998). Under TVAAS, students are measured each year, and data are compiled into a longitudinally merged database linking individual outcomes to teachers, schools, and districts (Chatterji 2002). Using a mixed model somewhat like the models estimated in this article, TVAAS produces estimates of achievement gains for each school and teacher and then determines a school's performance by comparing the school's or teacher's gains to the norm group's gain on a given grade-level test (Kupermintz 2002).

Tennessee is not the only state with learning-based accountability. North and South Carolina have implemented systems similar to TVAAS, as has the city of Dallas (Ladd and Walsh 2002). Since the NCLB was passed, politicians and lawmakers have also come to recognize the advantages of learning-based measures. Indeed, in 2005, the U.S. secretary of education announced that states could collect data on children's learning or achievement growth (along with current information on raw achievement) that will eventually be used for accountability purposes. Several states have since gained approval to pilot-test "growth model" accountability systems.[2] Outside the policy-making arena, scholars of education have produced a wide range of indictors of students' learning (for overviews, see Scheerens and Bosker 1997; Teddlie and Reynolds 1999).

Our extension of this useful work is to note an important limitation to learning-based measures of school effectiveness—the amount learned in a year is still heavily influenced by children's time outside school. The simplest way to understand how schools lack control over students' learning is to recognize that even during the academic year, children spend most of their time *outside the school environment*. Table 1 presents calculations for the proportion of waking hours spent in school, estimated for students with perfect attendance. During a calendar year, which includes the nonschool summer, the proportion is .25. If we focus on the academic year only, the proportion of time spent in school increases, but only to .32. These calculations agree closely with the survey estimates of Hofferth and Sandberg (2001), who reported that school-age children are awake an average of 99–104 hours per week and spend 32–33 of these hours in school. In short, whether we measure children's gains over a calendar or academic year, the majority of children's waking hours are spent outside school. And if we include the years before kindergarten—which certainly affect achievement and may also affect later learning—we

find that the typical 18-year-old American has spent only 13 percent of his or her waking hours in school (Walberg 1984).

In short, even during the academic year, children spend most of their time outside school. As a result, through no special effort of their own, schools that serve children from advantaged nonschool environments will more easily register learning or gains than will schools that serve children from poor nonschool environments. Learning, then, although more under schools' control than achievement, is still heavily contaminated by nonschool factors.

Covariate Adjustment

One way to address the problem of nonschool influences is to adjust schools' learning rates or achievement levels statistically using measured characteristics of students or covariates. But this approach has serious problems (cf. Rubenstein et al. 2004).

First, as a practical matter, it is difficult to find well-measured covariates that account fully for children's nonschool environments. While past research has tried to account for nonschool differences using measures of poverty, race/ethnicity, and family structure, among other influences (Ladd and Walsh 2002), it is rarely clear whether a sufficient number of nonschool confounders have been measured and measured well (Meyer 1996). Even when considerable nonschool information is available, it may not adequately capture the effect of nonschool influences on learning. Typical measures of the nonschool environment, such as parents' socioeconomic status (SES), family structure, race/ethnicity, and gender, explain only 30 percent of the variation in children's cognitive skills at the beginning of kindergarten and just *1 percent* of the variation in the amount that children learn when they are out of school during summer vacation (Downey et al. 2004).

It is also possible for covariates to *overcorrect* estimates of school effectiveness. For example, suppose that students' race/ethnicity and SES are correlated with unmeasured variables that affect school quality. Models that remove the effects of race/ethnicity and SES may also remove the effect of the unmeasured school-quality variables. To take an extreme example, consider a segregated school system, in which white children and black children attend separate schools. By adjusting for students' race, one is saying, in effect, that an all-black school can be compared only to another all-black school. Under

Table 1. Proportion of Waking Hours that Children Spend in School

Hours	From Birth to Age 18	One Calendar Year	One Academic Year
Hours in school per day	—	7	7
School days attended per year	—	180	180
Hours awake each day	—	14	14
Hours in school per year	—	1,260	1,260
Hours awake per year	—	14 hours per day x 365 days = 5,110	14 hours per day x 285 days = 3,990
Proportion of waking hours in school	.13	1,260 hours per year/ 5,110 hours per year = .25	1,260 hours per year/ 3,990 hours per year = .32
Source	Walberg (1984)	Authors' calculations	Authors' calculations

such constraints, it is impossible to see whether all-black schools are, on average, more effective or less effective than are all-white schools.

Finally, even if available covariates had more desirable statistical properties, adjusting for covariates such as race is politically sensitive. The idea that schools that enroll minority students are held to lower standards is troubling on many levels. Indeed, some of the popularity of the TVAAS system may stem from Sanders's claim that learning rates do not need adjustment, since they are unrelated to race and SES (Sanders 1998; Sanders and Horn 1998; see also Ryan 2004). As we will show, this claim is incorrect (see also Downey et al. 2004; Kupermintz 2002), although it is true that disadvantage is much less correlated with learning rates than with achievement.

In short, using covariates to adjust estimates of school quality has both methodological and political limitations. Our alternative strategy, described next, draws on seasonal comparison techniques developed as a way to improve on covariate adjustment.

Impact

As we mentioned earlier, measured characteristics, such as race and SES, seem to be weak and indirect proxies for the nonschool factors that affect children's learning rates. We now introduce a more direct approach to removing nonschool factors from school evaluations—an approach that we call *impact.*

Conceptually, impact is the difference between the rate at which children learn in school and the rate at which they would learn if they were never enrolled in school. The never-enrolled learning rate is a counterfactual (e.g., Winship and Morgan 1999), which, as usual, cannot be observed directly. However, we can observe how quickly children learn when they are out of school during summer vacation. As a practical matter, then, we can estimate a school's impact by subtracting its students' summer learning rate from the students' school-year learning rate. For example, in this article, we define a school's impact as the average difference between its students' first-grade learning rate and its students' learning rate during the previous summer.

The idea of defining impact by comparing school learning rates to summer learning rates builds on Heyns's (1978) insight that while learning during the school year is a function of both nonschool and school factors, summer learning is a product of nonschool factors alone. By focusing on the degree to which schools increase children's learning over the rates that prevail when children are not in school, the impact measure aims to separate school effects from nonschool effects on learning.

A key advantage of the impact measure is that it circumvents the formidable task of trying to measure and statistically adjust for all the different aspects of children's nonschool environments. By focusing instead on nonschool learning, impact arguably captures what we need to know about children's learning opportunities outside school without incurring the methodological and political problems of covariate adjustment. Another advantage of the impact approach is that it does not assume that variations in learning rates are solely a function of *environmental* conditions. Even nonenvironmental effects on learning (e.g., potential variations in students' innate motivation levels) are better accounted for with summer–school-year comparisons.

An estimate of impact requires seasonal data—that is, achievement scores collected at both the beginning and end of successive school years. The notable advantage of seasonal data is that they provide an estimate of children's rate of cognitive growth during the summer, when children are not in school. Seasonal data are rare in educational research, but are highly revealing when they are collected. For example, previous researchers have noted that gaps in academic skills widen primarily during the summer, rather than during the school year, suggesting that schooling constrains the growth of inequality (Downey et al. 2004; Entwisle and Alexander 1992, 1994; Heyns 1978; Reardon 2003).

Knowing how fast children learn when they are exposed full time to their nonschool environment provides critical leverage for iso-

lating school effects. For this reason, the impact measure has important practical advantages over accountability approaches that require extensive measures of the quality of students' nonschool environments. Even in detailed social surveys like the one analyzed in this article, measures of the nonschool environment are imperfect and incomplete, and most school systems collect far less information than does a social survey. The advantage of the impact measure is that it reduces dependence on *observed* nonschool characteristics, instead relying on the summer learning rate, which is presumably affected not only by observed characteristics, but by unobserved and even *unobservable* nonschool influences.

METHODS

Data

We used the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K), a survey administered by the National Center for Education Statistics (NCES), U.S. Department of Education (NCES 2003). ECLS-K follows a multistage sampling design—first sampling geographic areas, then sampling schools within each area, and finally sampling children within each school. Children were tracked from the beginning of kindergarten in fall 1998 to the end of fifth grade in spring 2004. But only in the first two school years were seasonal data collected that can be used to estimate school-year and summer learning rates.

We evaluated schools using reading and math tests. The reading tests measure five levels of proficiency: (1) identifying upper- and lower-case letters of the alphabet by name, (2) identifying letters with sounds at the beginning of words, (3) identifying letters with sounds at the end of words, (4) recognizing common words by sight, and (5) reading words in context. Math skill is also gauged by five levels of proficiency: (1) identifying one-digit numerals, (2) recognizing a sequence of patterns, (3) predicting the next number in a sequence, (4) solving simple addition and subtraction problems, and (5) solving simple multiplication and division problems and recognizing more complex number patterns. We focus our presentation on the results for reading; results for mathematics, which are generally similar, are presented in Appendix A.

The tests followed a two-stage format that was designed to reduce ceiling and floor effects. In the first stage, children took a "routing test" containing items of a wide range of difficulty. In the second stage, children took a test containing questions of "appropriate difficulty," given the results of the routing test. Item response theory (IRT) was used to map children's answers onto a common 64-point scale for mathematics and a 92-point scale for reading. (The reading scale was originally 72 points, but was rescaled when questions were added after the kindergarten year.) Few scores were clustered near the top or bottom of the IRT scales, suggesting that ceiling and floor effects were successfully minimized. In addition, the IRT scales improved reliability by downweighting questions with poor discrimination or high "guessability" (Rock and Pollack 2002).

The reading and mathematics scales may be interpreted in terms of average first-grade learning rates. A single point on the reading scale, for example, is approximately the amount learned in two weeks of the first grade. We can support this statement by pointing out that, during the first grade, reading scores increase at an average rate of about 2.57 points per month,.

A total of 992 schools were visited for testing in the fall of kindergarten (Time 1), in the spring of kindergarten (Time 2), and in the spring of first grade (Time 4). Among these 992 schools, 309 were randomly selected for an extra test in the fall of first grade (Time 3). Only in those 309 schools could we estimate first-grade and summer learning rates. Since the summer learning rate is interpreted as a window into the nonschool environment, we excluded children who spent part or all of the summer in school—that is, children who attended summer school or schools that used year-round calendars. We also excluded children who transferred schools during the two-year observation period, since it would be difficult to know which school deserved credit

for these students' learning.[3] In the end, our analysis focused on 4,217 children in 287 schools. On average, 15 children were tested per school, but in individual schools as few as 1 or as many as 25 students were tested. The results were not appreciably different if we restricted the sample to schools with at least 15 tested students.

Multilevel Growth Model

We estimated schools' achievement, learning, and impact rankings using a multilevel growth model (Raudenbush and Bryk 2002). Specifically, we fit a three-level model in which test scores (Level 1) were nested within children and children (Level 2) were nested within schools (Level 3). This multilevel approach allowed us to estimate mean levels of achievement, learning, and impact, as well as school-, child-, and test-level variation.

If each child was tested on the first and last day of each school year, then learning could be estimated simply by subtracting successive test scores. In the ECLS-K, however, schools were visited on a staggered schedule, so that, depending on the school, fall and spring measurements could be taken anywhere from one to three months from the beginning or end of the school year. To compensate for the varied timing of achievement tests, our model adjusts for the time that the children spent in kindergarten, summer vacation, and the first grade at the time of each measurement.

More specifically, at Level 1, we modeled each test score Y_{tcs} as a linear function of the months that child c in school s had been exposed to kindergarten, summer, and first grade at the time of test t:[4]

$$Y_{tcs} = \alpha_{0cs} + \alpha_{1cs}\,\text{KINDERGARTEN}_{tcs} + \alpha_{2cs}\,\text{SUMMER}_{tcs} + \alpha_{3cs}\,\text{FIRST GRADE}_{tcs} + e_{tcs} \quad (1a),$$

where there are

t = 1, 2, 3, 4 measurement occasions between the start of kindergarten and the end of first grade, for

c = 1, . . . , 15 or so children in each of

s = 1, . . . , 287 schools.

The slopes α_{1cs}, α_{2cs}, and α_{3cs} represent monthly rates of learning during kindergarten, summer, and the first grade, and the intercept α_{0cs} represents the child's achievement level on the last day of the first grade.[5] This last-day achievement level is an extrapolation; it is not the same as the final test score because the final test was typically given one to three months before the end of the first grade. The residual term e_{tcs} is measurement error, or the difference between the test score Y_{tcs} and the child's true achievement level at the time of the test. The variance of the measurement error can be calculated from test-reliability estimates in Rock and Pollack (2002); Table 2 reports the error variance for reading and math tests on each of the four test occasions.

In vector form, the Level 1 equation can be written concisely as

$$Y_{tcs} = \textbf{EXPOSURES}_{tcs}\,\boldsymbol{\alpha}_{cs} + e_{tcs} \quad (1b),$$

where $\alpha_{cs} = [\alpha_{0cs}\ \alpha_{1cs}\ \alpha_{2cs}\ v_{3cs}]^T$ and $\textbf{EXPOSURES}_{tcs} = [1\ \text{KINDERGARTEN}_{tcs}\ \text{SUMMER}_{tcs}\ \text{FIRST GRADE}_{tcs}]$.

Then the Level 2 equation models child-level variation within each school:

$$\alpha_{cs} = \beta_s + \boldsymbol{a}_c \quad (2),$$

where $\beta_s = [\beta_{0s}\ \beta_{1s}\ \beta_{2s}\ \beta_{3s}]^T$ is the average achievement level and learning rates for school s, and $\boldsymbol{a}_c = [a_{0c}\ a_{1c}\ a_{2c}\ a_{3c}]^T$ is a *random effect* representing the amount that child c deviates from the average for school s.

Likewise, the Level 3 equation models school-level variation between one school and another:

$$\beta_s = \gamma_0 + \boldsymbol{b}_s \quad (3),$$

where $\gamma_0 = \gamma_{00}\ \gamma_{01}\ \gamma_{02}\ \gamma_{03}]^T$ is a *fixed effect* representing the grand average achievement level and learning rates across all schools, and $\boldsymbol{b}_s = [b_{0s}\ b_{1s}\ b_{2s}\ b_{3s}]^T$ is a school-level random effect representing the departure of school s from the grand average. The Level 2 and 3 random effects $\boldsymbol{a}_c$ and $\boldsymbol{b}_s$ are assumed to be uncorrelated with each other; $\boldsymbol{a}_c$ and $\boldsymbol{b}_s$ are multivariate normal variables with means of zero and unrestricted covariance matrices of $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$.

Table 2. Measurement Error Variance on Four Reading Tests and Four Mathematics Tests

	Reading			Mathematics		
Occasion (*t*)	Total Variance	Reliability	Measurement Error Variance	Total Variance	Reliability	Measurement Error Variance
1. Fall 1998	73.62	0.93	5.15	50.55	0.92	4.04
2. Spring 1999	117.72	0.95	5.89	76.39	0.94	4.58
3. Fall 1999	160.53	0.96	6.42	92.35	0.94	5.54
4. Spring 2000	200.79	0.97	6.02	90.25	0.94	5.42

Note: Reliabilities were calculated by Rock and Pollack (2002) using IRT. If the reliability is *r* and the total variance of a test is *Var* (Y_{sct}), then the measurement error variance is (1-*r*) *Var* (Y_{sct}). Note that the variance changes (though not by much) from one measurement occasion to the next. Our analyses account for this heterogeneity, but ignoring it would yield similar results.

The Level 3 model can be expanded to include a vector of school characteristics $\boldsymbol{X}_s$:

$$\beta_s = \gamma_0 + \gamma_{1+}\boldsymbol{X}_s + \boldsymbol{b}_s \quad (4),$$

where γ_1 is a coefficient matrix representing the fixed effects of the school characteristics in $\boldsymbol{X}_s$, including the school's location (urban, rural, or suburban), ethnic composition (percentage minority), poverty level (percentage of students receiving free or reduced-priced lunches), and sector (public, Catholic, non-Catholic religious, or secular private).

Equations 1, 2, and 4 can be combined to give a mixed-model equation:

$$Y_{tcs} = \textsc{exposures}_{tcs}\,(\gamma_0 + \gamma_1\,\boldsymbol{X}_s + \boldsymbol{b}_s + \boldsymbol{a}_c) + e_{tcs} \quad (5),$$

which shows how differences in school learning rates are modeled using interactions between school characteristics $\boldsymbol{X}_s$ and students' $\textsc{exposures}_{tcs}$ to kindergarten, summer, and first grade. This model has been used before (e.g., Downey et al. 2004). What is new in this article is the emphasis on two derived quantities:

1. The first derived quantity, *impact*, is the difference between the first-grade and summer learning rates. For school *s*, impact is $\beta_{4s} = \beta_{3s} - \beta_{2s}$.
2. The second derived quantity, *12-month learning*, is the average monthly learning rate over a period consisting of 2.4 months of summer, followed by 9.6 months of the first grade. For school *s*, 12-month learning is $\beta_{5s} = \frac{1}{12}(2.4\beta_{2s} + 9.6\beta_{3s})$.

Average values for impact and 12-month learning can be obtained from any software that estimates linear combinations of model parameters. The variances and correlations involving impact and 12-month learning were estimated through auxiliary calculations that are described in Appendix B.

MULTIPLE IMPUTATION

We compensated for missing values using a multiple-imputation strategy (Rubin 1987) that filled in each missing value with 10 plausible imputations. To ensure that the imputations accounted for correlations among tests on the same child, we formatted the data so that each child's test scores appeared on a single line alongside the other variables (Allison 2002). To account for the interactions in Equation 5, we multiplied the components of the interaction before imputation and imputed the resulting products like any other variable (Allison 2002; von Hippel, under review).[6] To account for the difference between child- and school-level variables, we first created a school-level file that included the school-level variables as well as school averages of the child and test variables. We imputed this school file 10 times and then merged the imputed school files back with the observed child and test data.

Although our imputation model included test scores, none of the imputed test scores was used in the analysis. Excluding imputations of the dependent variable is a strategy known as multiple imputation, then deletion (MID), which increases efficiency and reduces biases resulting from misspecification of the imputation model (von Hippel 2007).

Although we believe that our imputation strategy is sound, we recognize that alternatives are possible. It is reassuring to note that our effects are large and robust; we analyzed these data using a variety of imputation strategies without material effects on the results.

RESULTS

In this section, we compare school evaluation methods based on achievement, learning, and impact. We focus on the results for reading. The results for mathematics, which were generally similar, are presented in Appendix A.

Which Schools Are Failing?

Table 3 summarizes the distribution of achievement, learning, and impact across the sampled schools. At the end of the first grade, the average achievement level is 59.33 points (out of 92). Children reach this achievement

level by learning at an average rate of 1.70 points per month during kindergarten, losing 0.08 points per month during the summer, and then gaining 2.57 points per month during first grade. So school impact—the difference between first-grade and summer learning rates—has an average value of 2.64 points per month.[7] In addition, 12-month learning—the average learning rate over the 12-month period from the end of kindergarten to the end of first grade—is 2.57 points per month. Note that if we did not have seasonal data, we would have to use this 12-month, or calendar-year, learning rate instead of the 9-month learning rate measured during the school year.

Of primary interest are the levels of agreement between different methods of evaluating schools. If agreement is high, then the methods are more or less interchangeable, and it does not matter much whether we evaluate schools in terms of achievement, learning, or impact. If agreement is low, however, then it is vital to know which method is best, since ideas about which schools are failing (or succeeding) would depend strongly on the yardstick by which schools are evaluated.

One way to evaluate agreement is to look at the school-level correlations in the bottom half of Table 3. In general, achievement is moderately correlated with school-year and 12-month learning rates, but only weakly correlated with impact. For example, across schools, achievement (at the end of the first grade) has a .52 correlation with the first-grade learning rate (95 percent CI: .40 to .64), and a .58 correlation with the 12-month learning rate (95 percent CI: .48 to .69), but achievement has just a .16 correlation with impact (95 percent CI: –.04 to .36).

Although these correlations are suggestive, they are somewhat abstract. To make differences among the evaluation methods more concrete, let us suppose that every school were labeled as either "failing" or "successful." Of course, definitions of failure vary across states, complicating our attempt to address this issue with national data. A useful exercise with this data, however, is to suppose that a school is failing if it is in the bottom quintile on a given criterion. The question, then, is this: How often will a school from the bottom quintile on one criterion also be in the bottom quintile on another? For example, among schools with failing achievement levels, what percentage are failing with respect to learning or impact? This percentage can be obtained by transforming the correlations in Table 3.[8]

The estimated agreement percentages are shown in Table 4. Again, evaluations based on achievement agree only modestly with evaluations based on learning, and achievement agrees poorly with impact. Among schools in the bottom quintile for achievement, 49 percent (95 percent CI: 42 percent to 56 percent) are in the bottom quintile for 12-month learning, 45 percent (95 percent CI: 38 percent to 52 percent) are in the bottom quintile for first-grade learning, and a mere 26 percent (95 percent CI: 18 percent to 36 percent) are in the bottom quintile for impact. (The chance level of agreement would be 20 percent.) There are also substantial disagreements between impact and learning; for example, among schools in the bottom quintile for impact, only 56 percent (95 percent CI: 48 percent to 64 percent) are in the bottom quintile for first-grade learning, and only 38 percent (95 percent CI: 29 percent to 48 percent) are in the bottom quintile for 12-month learning.

To illustrate the disagreements among evaluation methods, Figure 1 plots empirical Bayes estimates (Raudenbush and Bryk 2002) of achievement, learning, and impact for the 287 schools in our sample. The plots show concretely how schools with failing achievement levels are often not failing with respect to learning rates and may even be above average with respect to impact. Conversely, a few schools that are succeeding with respect to achievement appear to be failing with respect to learning or impact.

What Kinds *of Schools Are Failing?*

What are the outward characteristics of low-performing schools? Conventional wisdom suggests that failing schools tend to be urban public schools that serve predominantly poor or minority students. But conventional wis-

Table 3. Reading Achievement, Learning, and Impact, as Measured on a 92-point Scale

	Achievement,	Monthly Learning Rates				
	End of First Grade	Kindergarten	Summer	First Grade	12 Months[a]	Impact[b]
Fixed Effects (means)	59.33***	1.70***	-0.08	2.57***	1.99***	2.64***
	(58.40,60.26)	(1.64,1.76)	(-0.18,0.03)	(2.50,2.63)	(1.94,2.04)	(2.51,2.78)
Random Effects (school level)						
SD	7.07***	0.39***	0.57***	0.45***	0.36***	0.78***
	(6.32,7.81)	(0.33,0.44)	(0.46,0.69)	(0.40,0.51)	(0.32,0.40)	(0.63,0.93)
Correlations						
Kindergarten learning	0.40***					
	(0.25, 0.54)					
Summer learning	0.19†	-0.30**				
	(-0.02,0.40)	(-0.52,-0.09)				
First-grade learning	0.52***	-0.19*	-0.14			
	(0.40,0.64)	(-0.37,-0.01)	(-0.36,0.08)			
12-month learning	0.58***	-0.29***	0.21†	0.94***		
	(0.48,0.69)	(-0.46,-0.13)	(-0.01,0.42)	(0.91,0.97)		
Impact	0.16	0.11	0.82***	0.68***	0.39***	
	(-0.04,0.36)	(-0.11,0.33)	(-0.90,-0.74)	(0.57,0.80)	(0.20,0.58)	

Note: Child-level random effects are not shown. The 95 percent confidence intervals are in parentheses.

[a]Twelve-month learning is reckoned from the end of kindergarten to the end of first grade.

[b]Impact is the difference between the first-grade and summer learning rates.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table 4. Percentage of Schools from the Bottom (or Top) Quintile on Criterion A that Are in the Bottom (or top) Quintile for Criterion B

	Achievement,	Learning Rates (points per month)			
Leaning or Impact	End of First Grade	Kindergarten	Summer	First Grade	12 Months
Kindergarten learning	38 (31, 46)				
Summer learning	28 (19, 39)	10 (4, 17)			
First grade learning	45 (38, 52)	13 (7, 19)	14 (7, 23)		
12-month learning	49 (42, 56)	10 (5, 15)	28 (20, 38)	80 (76, 85)	
Impact	26 (18, 36)	24 (15, 34)	0 (0, 0)	56 (48, 64)	38 (29, 48)

Note: The 95% confidence intervals are in parentheses. These agreement rates can be derived from the correlations in Table 1 if it is assumed that school-level achievement, learning, and impact have an approximately normal distribution (as they appear to). See Appendix A for details.

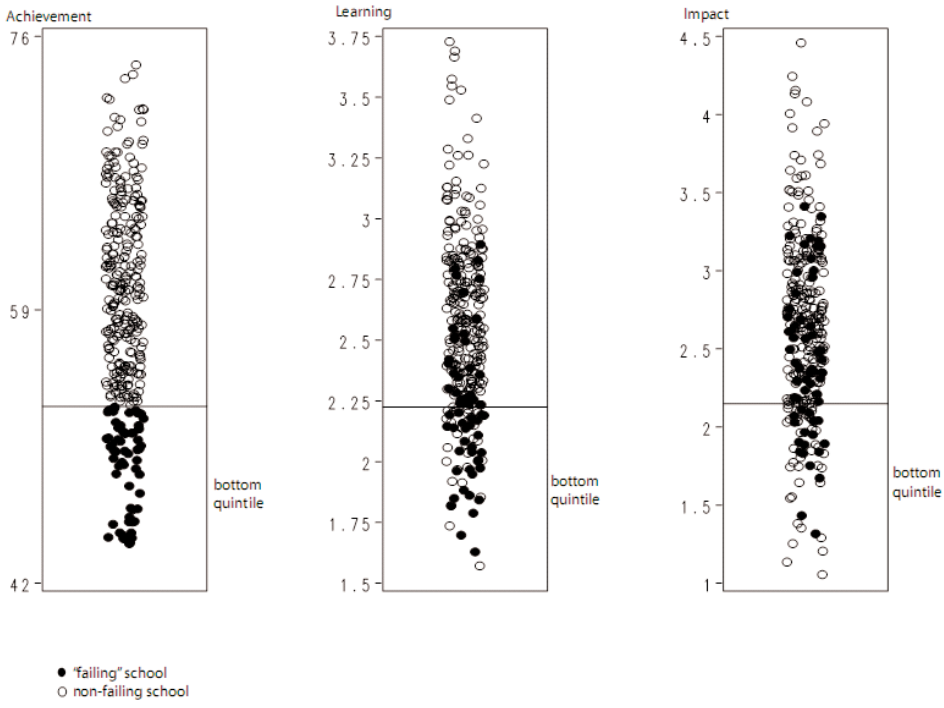

Figure 1. Schools That Are Failing with Respect to Achievement May Not Be Failing with Respect to Learning or Impact. (Only the vertical positions are meaningful; points have been horizontally dithered to reduce overplotting.)

dom is typically based on achievement scores. How might notions of school performance be challenged if schools were evaluated in terms of learning or impact? Table 5 presents the average characteristics of schools from the bottom and top four quintiles on empirical Bayes estimates of achievement, learning, and impact. The first column focuses on end-of-first-grade achievement levels. Here the results fit the familiar pattern. Compared to other schools, schools from the bottom achievement quintile tend to be public, rather than private, and urban, rather than suburban or rural. In addition, the students who attend schools from the bottom achievement quintile are more than twice as likely to come from minority groups and to qualify for free lunch programs.

When we evaluate schools on learning, however, socioeconomic differences between failing and successful schools shrink or even disappear. For example, when schools are categorized on the basis of first-grade learning, kindergarten learning, or 12-month learning, schools from the bottom quintile are not significantly more likely to be urban or public than are schools from the top four quintiles. Students at schools that rank in the bottom quintile for learning are more likely to be poor and minority than are students in the top four quintiles, but the ethnic and poverty differences when schools are evaluated on learning are at least 10 percent smaller than they are when schools are evaluated on achievement. When schools are evaluated on kindergarten learning, most of the socioeconomic differences between bottom-quintile schools and other schools are not statistically significant.

When schools are evaluated with respect to *impact*, the association between school characteristics and school failure is also weak—weaker than it is for first-grade or 12-month learning and almost as weak as it is for kindergarten learning. Under an impact criterion, schools from the bottom quintile are not significantly more likely to be urban or public than are schools from the top four quintiles, and low-impact schools do not have a disproportionate percentage of students who qualify for free lunches. Low-impact schools do

Table 5. Mean Characteristics of Failing versus Nonfailing Schools, Under Different Criteria for Failure

Characteristic	Achievement, End of First Grade			Kindergarten Learning			First-Grade Learning			12-Month Learning			Impact		
	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff
School Sector (Percentage of Schools)															
Public	96	71	***	75	76		79	75		79	75		69	78	
Catholic	4	12	†	14	10		5	12		5	12		11	10	
Other religious	0	12	**	8	10		6	11		7	10		7	10	
Secular private	1	5		3	4		9	3	*	9	3	*	12	2	***
School Location (Percentage of Schools)															
Urban	53	35	*	46	37		38	39		38	39		41	38	
Suburban	23	42	*	36	39		35	39		35	39		41	38	
Rural	24	23		18	24		27	22		27	22		18	24	
Proportion															
Receiving free lunches	52	22	***	34	27	†	38	26	**	38	26	**	26	29	
Receiving reduced-price lunches	9	7	†	8	7		8	7		8	7		7	7	
Minority	69	31	***	42	38		58	34	***	57	34	***	49	36	*

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

have a higher percentage of students from minority groups (49 percent versus 36 percent for the top four quintiles, $p < .05$), but the difference is about 10 percent smaller than it is when schools are evaluated on first-grade learning or 12-month learning and 25 percent smaller than it is when schools are evaluated on achievement.

Another way to examine school characteristics is to add school-level regressors to our multilevel model of achievement, learning, and impact. We do so in Table 6, which again shows that student disadvantage is more strongly associated with achievement than it is with learning or impact. Specifically, Table 6 indicates that school sector, school location, students' poverty, and minority enrollment explain 51 percent of the school-level variance in end-of-first-grade achievement levels but just 26 percent of the school-level variance in 12-month learning rates, 17 percent of the school-level variance in first-grade learning rates, 7 percent of the school-level variance in impact, and a mere 5 percent of the school-level variance in kindergarten learning rates. In short, when schools are evaluated with respect to achievement, schools that serve disadvantaged students are disproportionately likely to be labeled as failing. When schools are evaluated in terms of learning or impact, however, school failure appears to be less concentrated among disadvantaged groups.

REMAINING ISSUES

We have introduced impact as a potential replacement for the typically used achievement-based measures of school effectiveness. Yet we recognize that evaluating schools via impact requires some assumptions and raises several new questions.

First, the impact measure assumes that there is little "spillover" between seasons—that is, that school characteristics do not have important influences on summer learning. The literature on spillover effects is limited, but the available evidence suggests that spillover effects are minimal. Georgies (2003) reported no relationship between summer learning and kindergarten teachers' practices or classroom characteristics. And in our own supplemental analyses of ECLS-K, we found that summer learning rates were not higher if kindergarten teachers assigned summer book lists or if schools sent home preparatory "packages" before the beginning of the first grade.

Second, the impact measure assumes that nonschool influences on learning are similar during the school year and during summer vacation. This assumption is more debatable. It seems plausible that nonschool effects may be smaller during the school year than during the summer, for the obvious reason that during the school year, children spend less time in their nonschool environments. This observation suggests the possibility of a weighted impact score that subtracts only a fraction of the summer learning rate. The ideal weight to give summer is hard to know,[9] but the results for weighted impact would lie somewhere between the results for unweighted impact (effectively weighted impact where summer learning has a weight of one) and the results for school-year learning (which effectively gives summer learning a weight of zero). No matter where the results fall on this continuum, the characteristics of low-impact schools would be quite different from those of low-achieving schools. That is, compared to low-achievement schools, low-impact schools are not nearly as likely to be public, urban, poor, or heavily minority.[10]

An additional concern is that even if impact is a more *valid* measure of effectiveness than is achievement, it may also be less *reliable*. It is well known that estimates of school learning rates are less reliable than are estimates of school achievement levels (e.g., Kane and Staiger 2002; von Hippel 2004), and estimates of impact are less reliable still (von Hippel, forthcoming). In a companion paper, however, we show that the increase in validity more than compensates for the loss in reliability; that is, a noisy measure of learning is still a better reflection of school effectiveness than is a clean measure of achievement, and a noisy measure of impact may be better still (von Hippel forthcoming).

A final concern is that impact-based evaluation may penalize schools with high achievement. It may be difficult for any school, no

Table 6. School-level Predictors of Reading Achievement, Learning, and Impact

	Achievement,	Learning Rates (points per month)				Impact[b]
	End of First Grade	Kindergarten	Summer	First Grade	12 Months[a]	
Fixed Effects						
Intercept	64.16***	1.69***	–0.02	2.74***	2.14***	2.76***
	(62.24,66.08)	(1.55,1.84)	(–0.31,0.27)	(2.56,2.92)	(2.01,2.27)	(2.35,3.16)
School Sector						
Catholic	1.80	–0.03	0.07	0.09	0.08	0.02
	(–0.79,4.39)	(–0.23,0.17)	(–0.29,0.44)	(–0.14,0.31)	(–0.09,0.26)	(–0.46,0.50)
Other religious	4.39**	0.24*	0.37	0.13	0.18*	–0.24
	(1.47,7.30)	(0.02,0.46)	(–0.14,0.87)	(–0.12,0.37)	(0.00,0.36)	(–0.89,0.41)
Secular private	3.57	0.04	0.32	–0.52*	–0.34*	–0.85†
	(–1.04,8.18)	(–0.31,0.38)	(–0.40,1.05)	(–0.92,–0.12)	(–0.65,–0.03)	(–1.76,0.07)
School Location						
Urban	–0.41	–0.01	–0.07	0.00	–0.02	0.07
	(–2.22,1.40)	(–0.15,0.13)	(–0.33,0.19)	(–0.14,0.14)	(–0.13,0.09)	(–0.27,0.40)
Rural	–2.02*	0.11	–0.12	–0.05	–0.06	0.07
	(–3.95,–0.10)	(–0.05,0.26)	(–0.39,0.14)	(–0.21,0.12)	(–0.19,0.06)	(–0.28,0.42)
Proportion						
Receiving free lunches	–9.84***	–0.24	–0.52†	0.01	–0.10	0.53
	(–14.32,–5.36)	(–0.59,0.11)	(–1.14,0.10)	(–0.31,0.33)	(–0.35,0.14)	(–0.25,1.31)
Receiving reduced-priced lunches	–0.71	0.71	0.99	–0.29	–0.02	–1.28
	(–18.39,16.97)	(–0.62,2.04)	(–1.81,3.78)	(–1.98,1.39)	(–1.20,1.17)	(–5.13,2.56)
Minority	–5.39**	–0.03	0.08	–0.38**	–0.28***	–0.46
	(–8.73,–2.05)	(–0.28,0.21)	(–0.39,0.55)	(–0.63,–0.14)	(–0.45,–0.12)	(–1.10,0.17)
Random Effects (school level)						
SD	4.92***	0.37***	0.55***	0.41***	0.31***	0.76***
	(4.27,5.57)	(0.31,0.42)	(0.44,0.66)	(0.35,0.46)	(0.27,0.36)	(0.62,0.90)
R^2	.51	.05	.10	.17	.26	.07

Note: R^2 is the proportion by which the school-level variances are reduced from Table 1. Not shown: child- and test-level random effects, school-level correlations.

[a]The omitted school sector is public, and the omitted school location is suburban. Thus, the intercept represents the expected values for a suburban public school with no minority students and no students receiving free or reduced-price lunches.

†$p<.10$, *$p<.05$, **$p<.01$, ***$p<.001$. Parentheses enclose 95 confidence intervals.

matter how good, to accelerate learning during the school year for high-achievement children. Our study, however, did not find a negative correlation between impact and achievement; to the contrary, the correlation between achievement and impact was positive, although small (see Table 3). And among schools in the top quintile on achievement, 26 percent were also in the top quintile on impact (see Table 4), suggesting that it is possible for a high-achieving school to have high impact as well. It would be fair to say, though, that a school with fast summer learning cannot also have high impact (see Tables 4 and A2). To the degree that fast summer learning is typical of children from privileged families, this could be a concern for schools that serve such children under an impact-based evaluation system.

Although the assumptions of impact-based evaluation are nontrivial, we should bear in mind that *every* school-evaluation measure makes assumptions. The assumptions that are needed for the impact measure should be compared to those that are required to treat achievement or learning as measures of school performance. As we previously noted, evaluation systems that are based on achievement or learning assume that nonschool factors play a relatively minor role in shaping students' outcomes. This assumption is severely wrong for achievement and somewhat wrong for learning.

DISCUSSION

The simple observation that children are influenced in important ways by their nonschool environments undermines achievement-based methods for evaluating schools. Confidently identifying failing schools requires a method of evaluation that is sociologically informed—that is, a method that recognizes that children's cognitive development is a function of exposure to multiple social contexts. While holding schools accountable for their performance is attractive for many reasons, schools cannot reasonably be held responsible for what happens to children when they are not under their purview.

Other scholars have made the same observation and have proposed alternatives to achievement-based assessment by using annual learning rates or by "adjusting" achievement levels for schools' socioeconomic characteristics. We have already discussed the practical, theoretical, and political difficulties of these alternatives. Our contribution is a novel solution. By using seasonal data, we can evaluate schools in terms of impact—separating the effects of the school and nonschool environments without having to measure either environment directly. We suggest that impact can be an important part of the continuing discussion on measuring the effectiveness of schools.

We have argued that there are conceptual reasons for preferring impact over achievement and even over learning-based measures of school effectiveness. If we are correct that achievement is the least valid measure of school effectiveness, then our results suggest that there is substantial error in the way schools are currently evaluated. Indeed, our analyses indicate that, more often than not, schools that are vulnerable to the "failing" label under achievement standards are not among the least effective schools. Specifically, among schools from the bottom quintile for achievement, we found that less than half are in the bottom quintile for learning and only a quarter are in the bottom quintile for impact. In these mislabeled schools, students have low achievement levels, but they are learning at a reasonable rate, and they are learning substantially faster during the school year than during summer vacation. These patterns suggest that many so-called failing schools are having at least as much impact on their students' learning rates as are schools with much higher achievement scores.

We should emphasize that our results do not suggest that all schools have a similar impact. To the contrary, impact varies even more across schools than does achievement or learning. For impact, the between-school coefficient of variation is 30 percent; that is, the between-school standard deviation is 30 percent of the mean. For learning rates, by contrast, the coefficient of variation is just 23 percent in kindergarten and 18 percent in the first grade, and for end-of-first-grade achievement, the coefficient of variation is just 12

percent. So schools do vary substantially in impact, but variation in impact is not strongly associated with school characteristics, such as sector or location, or with the characteristics of the student body. Whereas high-achieving schools are concentrated among the affluent, high-impact schools exist in communities of every kind. For example, in disadvantaged communities, despite scarce resources, the high turnover of teachers, and low parental involvement, a sizable number of schools are having a considerable impact—much more than was previously thought. When we measure school effectiveness fairly, the results highlight how schools that serve the disadvantaged can do a good job even if they do not raise students' skills to a high or even average level of proficiency.[11]

Our results raise serious concerns about the current methods that are used to hold schools accountable for their students' achievement levels. Because achievement-based evaluation is biased against schools that serve the disadvantaged, evaluating schools on the basis of achievement may actually undermine the NCLB goal of reducing racial/ethnic and socioeconomic gaps in performance. If schools that serve the disadvantaged are evaluated on a biased scale, their teachers and administrators may respond like workers in other industries when they are evaluated unfairly—with frustration, reduced effort, and attrition (Hodson 2001). Under a fair system, a school's chances of receiving a high mark should not depend on the kinds of students the school happens to serve.

The validity of school performance measures is critical to the success of market-based educational reforms because making information about school quality publicly available is supposed to pressure school personnel to improve. But our results suggest that the information that is currently available regarding school quality is substantially flawed, undermining the development of market pressures as a mechanism for improving American schools. Achievement-based indicators of school effectiveness reduce market efficiency by too often sending parents away from good schools that serve children from disadvantaged backgrounds and insufficiently pressuring unproductive schools that serve children from advantaged backgrounds. Our results suggest that the magnitude of the error is substantial; indeed, current accountability systems that rely on achievement may do as much to undermine school quality as they do to promote it.

NOTES

1. In Ohio, adequate yearly progress typically means reducing the gap between a school's or district's baseline performance (average of the years 1999–2000, 2000–01, and 2001–02) and the proficiency bar by 10 percentage points per year between 2003–04 and 2013–14.

2. Pilot programs were first approved in 2006 by Tennessee and North Carolina and in 2007 by Arizona, Arkansas, Delaware, Florida, Iowa, and conditionally in Ohio.

3. A multilevel model requires that each unit from the lower level (each child) remains nested within a single unit from the higher level (a school). Data that violate this assumption may be modeled using a cross-classified model, but such models present serious computational difficulties, especially when a child's new school was not in the original sample. In our analyses, we deleted tests taken after a child moved schools. The results are not materially different if we keep these scores and attribute them to the child's original school.

4. These exposures are estimated by comparing the test date to the first and last dates of kindergarten and the first grade. Test dates are part of the public data release; the first and last dates of the school year are available to researchers with a restricted-use data license.

5. To ensure that the intercept had this interpretation, we centered each EXPOSURES variable around its maximum. To understand maximum centering, let $\text{KINDERGARTEN}^*_{tcs}$ be the number of months that child c in school s has spent in kindergarten at the time of test t. The maximum value of $\text{KINDERGARTEN}^*_{tcs}$ is KINDLENGTH_s, which is the length of the kindergarten year in school s. (An average value

would be KINDLENGTH$_s$ = 9.4 months.) Then the maximum-centered variable KINDERGARTEN$_{tcs}$ is defined as KINDERGARTEN*$_{tcs}$ – KINDLENGTH$_s$; this maximum-centered variable has a maximum of zero. If KINDERGARTEN$_{tcs}$, SUMMER$_{tcs}$ and FIRST GRADE$_{tcs}$ are all maximum centered, the intercept σ_{0cs} represents the child's score on the last day of the first grade, when KINDERGARTEN$_{tcs}$, SUMMER$_{tcs}$, and FIRST GRADE$_{tcs}$ all reach their maximum values of zero.

6. As is often the case, there was substantial colinearity between the interactions and the component variables. The imputation model compensated for this colinearity by using a ridge prior, as suggested by Schafer (1997).

7. 2.57 minus -0.08 gives an impact of 2.65, but if values are not rounded before subtraction, the value of impact is closer to 2.64.

8. The resulting percentages will be measures of latent school-level agreement, discounting random variation at the child and test levels. The transformation assumes that the different measures of school effectiveness have a multivariate normal distribution. (Scatterplots suggest that this assumption is reasonable.) Let (Z_i, Z_j) be standardized versions of two school-effectiveness measures, and let $q \approx -.84$ be the first quintile of the standard normal distribution. Then, given that Z_i is in the bottom quintile (i.e., $Z_s < q$). the probability that Z_j is also in the bottom quintile is $p_{ij} = P(Z_i<q \mid Z_j<q) = 5\ P(Z_i<q, Z_j<q) = 5\ \Phi_2(q, q, \rho_{ij})$, where $\Phi_2(q, q, \rho_{ij})$ is the bivariate cumulative standard normal density with correlation ρ_{ij}, evaluated at (q, q) (Rose and Smith 2002). A confidence interval for p_{ij} is obtained by transforming the endpoints of a confidence interval for ρ_{ij}.

9. An initially attractive possibility is to estimate the fraction by regressing the school-year learning rate on the summer learning rate. But since the correlation between school-year and summer learning is *negative* (see Table 3), the estimated fraction would be negative as well, yielding an impact measure that is the sum, rather than the difference, of school and summer learning rates.

10. A more subtle possibility is that the nonschool effect on learning varies across seasons *and* the seasonal pattern varies across schools that serve different types of students. Suppose high-SES parents, for example, invest substantially in the summer but then relatively less so during the school year while low-SES parents produce the opposite seasonal pattern. This kind of scenario would produce biases in the impact measure, underestimating school impact for schools serving high-SES families and overestimating the performance of schools serving low-SES parents. Although little is known about this possible source of bias, most of what we know about parental involvement in children's schooling suggests that this pattern is unlikely. Socioeconomically advantaged parents remain actively involved in their children's lives during the academic year by helping with homework, volunteering in classes, and attending school activities and parent-teacher conferences (Lareau 2000).

11. It is not impractical to make school evaluation systems fairer. Currently, the NCLB requires testing at the end of every year in Grades 3–8. These tests are typically used to rank schools on the basis of achievement, but the availability of annual test scores makes it possible to rank schools on the basis of the amount learned in a 12-month calendar year. Alternatively, the six tests given in Grades 3–8 could be rescheduled to permit seasonal comparisons. The testing schedule could be reshuffled so that tests are given at the end of Grade 3 and the beginning and end of Grade 4 and then, likewise, at the end of the Grade 7 and the beginning and end of Grade 8. Such a schedule would allow school evaluators to estimate impact and school-year learning during the fourth grade and the eighth grade without increasing the number of tests. Valid information about these two school years would be preferable to the six years of low-validity achievement levels that are currently provided.

APPENDIX A

Results for Mathematics

Table A1. Mathematics Achievement, Learning, and Impact, as Measured on a 64-point Scale

	Achievement,	Monthly Learning Rate				Impact
	End of First Grade	Kindergarten	Summer	First Grade	12 Months	
Fixed Effects (means)	45.58***	1.34***	0.47***	1.57***	1.33***	1.10***
	(45.01,46.15)	(1.30,1.399)	(0.37,0.57)	(1.53,1.61)	(1.30,1.36)	(0.98,1.22)
Random Effects						
School-level *SD*	4.26***	0.27***	0.58***	0.26***	0.20***	0.70***
	(3.79,4.72)	(0.24,0.31)	(0.48,0.68)	(0.22,0.30)	(0.17,0.23)	(0.58,0.83)
School-level correlations						
Kindergarten learning	0.44***					
	(0.29,0.59)					
Summer learning	0.07	–0.44***				
	(–0.12,0.27)	(–0.62,–0.26)				
First-grade learning	0.11	–0.22*	–0.31**			
	(-0.06,0.28)	(–0.42,–0.03)	(–0.51,–0.12)			
12-month learning	0.15†	–0.50***	0.30**	0.81***		
	(–0.01,0.32)	(–0.65,–0.34)	(0.11,0.50)	(0.73,0.88)		
Impact	–0.02	0.28**	–0.94***	0.63***	0.05	
	(–0.21,0.17)	(0.07,0.48)	(–0.96,–0.91)	(0.49,0.76)	(–0.17,0.26)	

Note: Parentheses enclose 95 percent confidence intervals. Random effects at the child level are not shown.
[a]Twelve-month learning is reckoned from the end of kindergarten to the end of first grade.
[b]Impact is the difference between the first-grade learning rate and the summer learning rate.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table A2. Which Schools Are Failing Mathematics? Percentage Agreement Matrix

	Achievement, End of First Grade	Learning Rates (points per month)			
		Kindergarten	Summer	First Grade	12 Months
Kindergarten learning	40 (33,49)				
Summer learning	23 (16,31)	6 (2,11)			
First-grade learning	24 (18,32)	12 (6,19)	9 (4,16)		
12-month learning	26 (20,34)	4 (1,8)	33 (24,44)	65 (59,73)	
Impact	19 (12,27)	32 (23,42)	0 (0,0)	51 (43,61)	22 (14,31)

Note: Among schools from the bottom quintile on criterion A, this matrix shows what percentage are also in the bottom quintile for criterion B. Parentheses enclose 95 percent confidence intervals.

Table A3. Mean Characteristics of Failing versus Nonfailing Schools, Under Different Criteria for Failure

Characteristic	Achievement, End of First Grade			Kindergarten Learning			First-Grade Learning			12-Month Learning			Impact		
	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff	Bottom Quintile	Top Four Quintiles	Diff
School Sector (percentage of schools)															
Public	98	70	***	86	73	*	67	78	†	65	79	*	72	77	
Catholic	0	13	**	2	13	*	15	9		14	10		14	10	
Other religious	2	12	*	6	11		8	10		10	10		6	11	
Secular private	0	5	†	5	3		11	2	**	11	2	**	8	3	
School Location (percentage of schools)															
Urban	54	35	**	51	36	*	40	38		39	39		37	39	
Suburban	16	44	***	33	40		38	38		38	38		46	36	
Rural	30	21		16	25		22	23		23	23		16	25	
Average Percentage of Students															
Receiving free lunches	52	22	***	43	24	***	27	28		26	29		25	29	
Receiving reduced-price lunches	9	7	*	8	7	†	7	7		7	7		7	7	
Minority	73	30	***	59	33	***	42	38		42	38		39	39	

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Table A4. School–level Predictors of Mathematics Achievement, Learning, and Impact

	Achievement,	Learning Rates (points per month)				Impact[b]
	End of First Grade	Kindergarten	Summer	First Grade	12 Months[a]	
Fixed Effects						
Intercept	49.72***	1.37***	0.60***	1.62***	1.40***	1.02***
	(48.67,50.78)	(1.27,1.48)	(0.36,0.84)	(1.52,1.72)	(1.32,1.48)	(0.74,1.30)
School Sector						
Catholic	0.60	0.06	0.13	–0.15*	–0.09	–0.28
	(–0.91,2.10)	(–0.08,0.21)	(–0.22,0.49)	(–0.28,–0.02)	(–0.19,0.02)	(–0.70,0.14)
Other religious	0.59	0.09	–0.13	–0.03	–0.05	0.09
	(–1.17,2.36)	(–0.07,0.25)	(–0.55,0.29)	(–0.20,0.13)	(–0.17,0.06)	(–0.43,0.61)
Secular private	–0.82	–0.24†	0.17	–0.44**	–0.30**	–0.60
	(–3.50,1.86)	(–0.50,0.02)	(–0.50,0.83)	(–0.70,–0.18)	(–0.51,–0.10)	(–1.40,0.19)
School Location						
Urban	0.17	0.01	–0.21	0.06	0.01	0.27
	(–0.90,1.23)	(–0.10,0.11)	(–0.49,0.08)	(–0.03,0.16)	(–0.06,0.07)	(–0.07,0.62)
Rural	–1.47*	0.08	–0.24†	0.01	–0.04	0.25
	(–2.61,–0.33)	(–0.03,0.19)	(–0.52,0.04)	(–0.10,0.13)	(–0.12,0.04)	(–0.09,0.60)
Proportion						
Receiving free lunches	–5.61***	–0.15	–0.23	0.08	0.01	0.32
	(–8.33,–2.89)	(–0.41,0.10)	(–0.78,0.31)	(–0.13,0.29)	(–0.15,0.18)	(–0.34,0.97)
Receiving reduced-price lunches	–4.27	0.42	0.95	–0.44	–0.14	–1.39
	(–14.15,5.62)	(–0.51,1.34)	(–1.10,3.01)	(–1.44,0.56)	(–0.89,0.61)	(–3.97,1.18)
Minority	–5.37***	–0.11	0.00	–0.10	–0.08	–0.10
	(–7.08,–3.65)	(–0.27,0.05)	(–0.38,0.38)	(–0.24,0.04)	(–0.18,0.03)	(–0.55,0.36)
Random Effects						
School-level *SD*	2.65***	0.26***	0.56***	0.24***	0.19***	0.67***
	(2.24,3.06)	(0.22,0.29)	(0.46,0.66)	(0.20,0.28)	(0.16,0.22)	(0.55,0.79)
R^2	.61	.07	.07	.15	.10	.08

Note: R^2 is the proportion by which the school-level variances are reduced from Table A1. Child- and test-level random effects and school-level correlations are not shown.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

APPENDIX B

STATISTICAL METHODS

Our basic multilevel growth model estimates end-of-first-grade achievement levels as well as learning rates during kindergarten, summer, and the first grade. But transforming these quantities into impact and 12-month learning requires some extra calculation.

More specifically, if $\boldsymbol{\beta}_s = [\beta_{0s}\ \beta_{1s}\ \beta_{2s}\ \beta_{3s}]^T$ represents the average end-of-first grade achievement level and the average kindergarten, summer, and first-grade learning rates for school *s*, then the school-level equation is

$$\boldsymbol{\beta}_S = \gamma_0 + \gamma_1 \boldsymbol{X}_s + \boldsymbol{b}_s \qquad \text{(A1)},$$

Where $\gamma_0 = [\gamma_{00}\ \gamma_{10}\ \gamma_{20}\ \gamma_{30}]^T$ is a fixed intercept, $\gamma_1 = \gamma_{10}\ \gamma_{11}\ \gamma_{12}\ \gamma_{13}]^T$ is a fixed matrix of slopes representing the effects of the school characteristics in $\boldsymbol{X}_s$, and $\boldsymbol{b}_s = [b_{0s}\ b_{1s}\ b_{2s}\ b_{3s}]^T$ is a school-level random effect with a mean of zero and a covariance matrix of $\boldsymbol{\Sigma}_b$. For certain purposes, it will be convenient to work with $vech(\boldsymbol{\Sigma}_b)$, which is a vector containing all the nonredundant elements of $\boldsymbol{\Sigma}_b$—the lower triangle of $\boldsymbol{\Sigma}_b$, beginning with the first column[a] (Harville 1997)

Multilevel modeling software (such as the MIXED procedure in SAS) provides point estimates $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\Sigma}_b$, as well as asymptotic covariance matrices $V(\hat{\gamma}_0)$, $V(\hat{\gamma}_1)$, and $V(vech(\hat{\Sigma}_b))$ that represent the uncertainty in the point estimates. (The diagonal elements of the asymptotic covariance matrices are squared standard errors.)

Combining these estimates to obtain estimates of 12-month learning and impact requires some transformation. As we noted in the text, the impact of school *s* is $\beta_{4s} = \beta_{3s} - \beta_{2s}$, or $\beta_{4s} = \boldsymbol{c}_{impact}\ \boldsymbol{\beta}_s$, where $\boldsymbol{c}_{impact} = [0\ 0\ -1\ 1]$. Likewise, the 12-month learning rate in school *s*—the average monthly learning rate over a 12-month period consisting (on average) of 2.4 months of summer followed by 9.6 months of first grade—is $\beta_{5s} = \frac{1}{12}(2.4\ \beta_{2s} + 9.6\beta_{3s}$ or, in vector form, $\beta_{5s} = \boldsymbol{c}_{12month}\ \beta_s$ where $\boldsymbol{c}_{12month} = \frac{1}{12}[0\ 0\ 2.4\ 9.6]$. So, if we let

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_4 \\ \mathbf{c}_{12month} \\ \mathbf{c}_{impact} \end{bmatrix}, \text{ where } \mathbf{I}_4 \text{ is the 4-by-4 identity matrix,} \qquad \text{(A2)},$$

then $\beta^*_s = \mathbf{C}\ \beta_s = [\beta_{0s}\ \beta_{1s}\ \beta_{2s}\ \beta_{3s}\ \beta_{4s}\ \beta_{5s}]^T$ is an expanded school-level vector that includes impact and 12-month learning as well as achievement, school-year, and summer learning. The following school-level equation represents how this expanded vector varies across schools:

$$\beta^*_s = \gamma^*_0 + \gamma^*_1 \boldsymbol{X}_s + \boldsymbol{b}^*_s \qquad \text{(A3)},$$

where $\gamma^*_0\ C\gamma_0$ and $\gamma^*_1 = C\gamma_1$ are the fixed intercept and slope, and the random effect $\boldsymbol{b}^*_s$ has a covariance matrix of $\Sigma^*_b = C\Sigma_b C^T$. Estimated parameters for this expanded equation (A3) can be derived from the estimates for the basic equation (A1), as follows: $\hat{\gamma}^*_0 = C\hat{\gamma}_0$, $\hat{\gamma}^*_1 = C\hat{\gamma}_1$ and $\hat{\Sigma}_b = C\hat{\Sigma}_b C^T$, or $vech(\hat{\Sigma}^*_b) = \mathbf{F}vech(\hat{\Sigma}_b)$, where $\mathbf{F} = \mathbf{H}_6\ (\mathbf{C}\otimes\mathbf{C})\mathbf{G}_4$, with $\mathbf{G}_4$ a duplication matrix and $\mathbf{H}_6$ an inverse duplication matrix.[b] The asymptotic covariance matrices for these transformed parameter estimates are $V(\hat{\gamma}^*_0) = \mathbf{C}V(\hat{\gamma}_0)\mathbf{C}^T$, $V(\hat{\gamma}^*_1) = \mathbf{C}V(\hat{\gamma}_1)\mathbf{C}^T$, and $V(vech(\hat{\Sigma}^*_b)) = \mathbf{F}V(vech(\hat{\Sigma}_b))\mathbf{F}^T$.[c]

The final step in our calculations is to convert the variances and covariances in Σ^*_b into standard deviations and correlations, which are easier to interpret. This is straightforward; a standard deviation σ is just the square root of the corresponding variance σ^2, and there is a simple matrix formula $\mathbf{R}^*_b = \mathbf{R}(\Sigma^*_b)$ for converting a covariance matrix such as Σ^*_b into a correlation matrix $\mathbf{R}^*_b$ (Johnson and Wichern, 1997). Again, it will be convenient to work with $vecp(\mathbf{R}^*_b)$, which is a vector containing the nonredundant elements of $\mathbf{R}^*_b$—the lower triangle of $\mathbf{R}^*_b$ excluding the diagonal, starting with the first column (Harville 1997).

Standard errors for the standard deviations and correlations that result from these calculations can be obtained using the delta rule (e.g., Agresti 2002, Sec. 14.1.3). For example, if $\hat{V}(\hat{\sigma}^2)$ is the squared standard error for the

variance estimate $\hat{\sigma}^2$, then $\hat{V}(\hat{\sigma}) = (\frac{d\hat{\sigma}}{d\hat{\sigma}^2})^2 \hat{V}(\hat{\sigma}^2) = \frac{1}{4\hat{\sigma}^2} \hat{V}(\hat{\sigma}^2)$ is the squared standard error for the standard deviation estimate $\hat{\sigma}$. Likewise, if $V(vech(\hat{\Sigma}^*_b))$ is the asymptotic covariance matrix of $vech(\hat{\Sigma}^*_b)$, then

$$V(vecp(\hat{\mathbf{R}}^*_b)) = \left[\frac{dvecp(\mathbf{R}(\hat{\Sigma}^*_b))}{dvech(\hat{\Sigma}^*_b)}\right] V(vech(\hat{\Sigma}^*_b)) \left[\frac{dvecp(\mathbf{R}(\hat{\Sigma}^*_b))}{dvech(\hat{\Sigma}^*_b)}\right]^{\mathrm{T}} \quad \text{(A4)}$$

is the asymptotic covariance matrix of $vecp(\mathbf{R}^*_b)$.

---

[a]In SAS software, the vector form of a symmetric matrix is called SYMSQR and begins with the first row, rather than the first column. The elements of SYMSQR(Σ_b) must be rearranged to obtain $vech(\Sigma_b)$.

[b]Duplication and inverse duplication matrices are defined in section 16.4b of Harville (1997). The relationship between $vech(\hat{\mathbf{\Sigma}}_a)$ and $vech(\mathbf{C}\hat{\mathbf{\Sigma}}_a\mathbf{C}^{\mathrm{T}})$ is given, using different notation, by formula 4.25 in Chapter 16. Formula 4.25 is restricted to the case where **C** is a square matrix; we use a generalization appropriate to the case where **C** is not square. We thank David Harville for suggesting this generalization (personal communication, September 27, 2005).

[c]These formulas make use of the general formula that if the vector $\boldsymbol{X}$ has mean μ and covariance matrix Σ, then the vector $\mathbf{A}\boldsymbol{X}$, where **A** is a matrix, has mean $\mathbf{A}\mu$ and covariance matrix $\mathbf{A}\Sigma\mathbf{A}^{\mathrm{T}}$ (Johnson and Wichern 1997:79).

REFERENCES

Agresti, Alan. 2002. *Categorical Data Analysis* (2nd ed.). New York: John Wiley.

Allison, Paul D. 2002. *Missing Data*. Thousand Oaks, CA: Sage.

Chatterji, Madhabi. 2002. "Models and Methods for Examining Standards-Based Reforms and Accountability Initiatives: Have the Tools of Inquiry Answered Pressing Questions on Improving Schools?" *Review of Educational Research* 72:345–86.

Chubb, John, and Terry Moe. 1990. *Politics, Markets, and America's Schools.* Washington, DC: Brookings Institution.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. Equality of Opportunity. Washington: Government Printing Office.

Downey, Douglas B., Paul T. von Hippel, and Beckett Broh. 2004. "Are Schools the Great Equalizer? Cognitive Ineqality during the Summer Months and the School Year." *American Sociological Review* 69:613–35.

Entwisle, Doris R. and Karl L. Alexander. 1992. "Summer Setback: Race, Poverty, School Composition and Math Achievement in the First Two Years of School." *American Sociological Review* 57:72–84.

—-. 1994. "The Gender Gap in Math: Its Possible Origins in Neighborhood Effects." *American Sociological Review* 59:822–38.

Georgies, Annie. 2003. "Explaining Divergence in Rates of Learning and Forgetting Among First Graders." Paper presented at the annual meeting of the American Sociological Association, Atlanta, GA.

Hart, Betty, and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brookes.

Harville, David. 1997. *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Heyns, Barbara. 1978. *Summer Learning and the Effects of Schooling*. New York: Academic Press.

Hodson, Randy. 2001. *Dignity at Work*. New York: Cambridge University Press.

Hofferth, Sandra L., and John F. Sandberg. 2001. "How American Children Spend Their Time." *Journal of Marriage and the Family*, 63:295–308.

Jencks, Christopher, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohen, Herbert Gintis, Barbara Heyns, and Stephan Michelson. 1972. *Inequality: A Reassessment of the Effect of Family and Schooling in America.* New York: Basic Books.

Johnson, Richard A., and Dean W. Wichern. 1997. *Applied Multivariate Statistical Analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Kane, Thomas J., and Douglas O. Staiger. 2002. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." Pp. 235–69 in *Brookings Papers on Education Policy*, edited by Diane Ravitch. Washington, DC: Brookings Institution.

Kupermintz, Haggai. 2002. "Value-Added Assessment of Teachers: The Empirical Evidence." Pp. 217–34 in *School Reform Proposals: The Research Evidence* edited by Alex Molnar. Greenwich, CT: Information Age Publishing.

Ladd, Helen. 2002. *Market-Based Reforms in Urban Education*. Washington, DC: Economic Policy Institute.

Ladd, Helen F., and Randall P. Walsh. 2002. "Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right." *Economics of Education Review* 21:1–27.

Lareau, Annette. 2000. *Home Advantage: Social Class and Parental Intervention in Elementary Education.* Oxford, England: Rowman & Littlefield.

Lee, Valerie E., and David T. Burkam. 2002. *Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School.* Washington, DC: Economic Policy Institute.

Meyer, Robert H. 1996. "Value-Added Indicators of School Performance." Pp. 197–223 in *Improving America's Schools: The Role of Incentives*, edited by Eric A. Hanushek and Dale W. Jorgenson. Washington, DC: National Academy Press.

National Center for Education Statistics. 2003. *Early Childhood Longitudinal Survey, Kindergarten Cohort*. Washington, DC: Author.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reardon, Sean. 2003. "Sources of Educational Inequality." Paper presented at the annual meeting of the American Sociological Association, Atlanta, GA.

Renzulli Linda A., and Vincent J. Roscigno. 2005. "Charter School Policy, Implementation, and Diffusion Across the United States." *Sociology of Education* 78:344–66.

Rock, Donald A., and Judith M. Pollack. 2002. *Early Childhood Longitudinal Study—Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade* (NCES 200205). Washington, DC: National Center for Education Statistics.

Rose, Colin, and Murray D. Smith. 2002. *Mathematical Statistics with Mathematica*. New York: Springer.

Rubenstein, Ross, Leanna Stiefel, Amy Ellen Schwartz, and Hella Bel Hadj Amor. 2004. "Distinguishing Good Schools from Bad in Principle and Practice: A Comparison of Four Methods." Pp. 55–70 in *Developments in School Finance: Fiscal Proceedings from the Annual State Data Conference of July 2003* (NCES 2004-325), edited by William J. Fowler, Jr. Washington, DC: U.S. Government Printing office.

Rubin, Donald B. 1987. *Multiple Imputation for Survey Nonresponse*. New York: John Wiley.

Ryan, James E. 2004. "The Perverse Incentives of the No Child Left Behind Act." *New York University Law Review* 79:932–89.

Sanders, William L. 1998. "Value-Added Assessment." *The School Administrator* 55(11):24–32.

Sanders, William L., and Sandra P. Horn. 1998. "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research." *Journal of Personnel Evaluation in Education* 12:247–56.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman and Hall.

Scheerens, Jaap, and Roel J. Bosker. 1997. *The Foundations of Educational Effectiveness*. Oxford, England: Pergamon Press.

Teddlie, Charles, and David Reynolds. 1999. *The International Handbook of School Effectiveness Research: An International Survey of Research on School Effectiveness*. London: Falmer Press.

Thernstrom, Abigail, and Stephan Thernstrom. 2003. *No Excuses: Closing the Racial Gap in Learning*. New York: Simon & Schuster.

von Hippel, Paul T. 2004. "School Accountability [a comment on Kane and Staiger (2002)]." *Journal of Economic Perspectives*, 18, 275–76.

—. 2007. "Regression with Missing Y's: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37, 83–117.

—. Forthcoming. "Achievement, Learning, and Seasonal Impact as Measures of School Effectiveness: It's Better to Be Valid than Reliable." *School Effectiveness and School Improvement*.

—. Under review. "Imputing Interactions."

Walberg, Herbert J. 1984. "Families as Partners in Educational Productivity." *Phi Delta Kappan* 65:397–400.

Wallberg, Herbert J. and Joseph L. Bast. 2003. Education and Capitalism: *How Overcoming Our Fear of Markets and Economics Can Improve America's Schools*. Stanford, CA: Hoover Institution Press.

West, Jerry, Kristin Denton, and Elvira Germino-Hausken. 2000. *America's Kindergartners: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99* (NCES 2000-070). Washington DC: National Center for Education Statistics.

Winship, Christopher, and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659–707.

Douglas B. Downey, Ph.D., *is Professor, Department of Sociology, The Ohio State University. His main fields of interest are stratification, race/ethnicity, and education. Dr. Downey is currently exploring when and why the black/white gap in cognitive skills emerges among young children (with Benjamin Gibbs) and is using seasonal comparisons to articulate how schooling influences social problems in the United States (with Paul T. von Hippel).*

Paul T. von Hippel, Ph.D., *is a doctoral candidate in sociology at the Ohio State University and a senior fraud analyst in the banking industry. His recent publications include a study of how schools restrain weight gain in early childhood* (American Journal of Public Health, *2007, with Brian Powell, Douglas B. Downey, and Nicholas Rowland) and an article on how to do regression when some values of the dependent variable are missing* (Sociological Methodology, *2007). One of his forthcoming articles (in* School Effectiveness and School Improvement) *explores the validity and reliability of the impact measure developed in this article.*

Melanie M. Hughes, Ph.D., *is Assistant Professor, Department of Sociology, University of Pittsburgh. Her main fields of interest are political sociology, gender, and stratification. Dr. Hughes is currently studying variation in the election of minority women around the world. In other pro-*

jects, she is examining the global spread of women's international nongovernmental organizations (with Pamela Paxton) and analyzing how wars affect women's political outcomes in Africa (with Aili Tripp).

This project was funded by grants from the NICHD (R03 HD043917-01), the Spencer Foundation (to Downey and von Hippel), the John Glenn Institute (to Downey), and the Ohio State P-12 Project (to Downey). The contributions of the first two authors are equal. We appreciate the comments of Beckett Broh, Donna Bobbitt-Zeher, Benjamin Gibbs, and Brian Powell. *Direct all correspondence to Douglas B. Downey, Department of Sociology, The Ohio State University, 300 Bricker Hall, 190 North Oval Mall, Columbus, OH 43210; e-mail: downey.32@osu.edu.*