

Are “Failing” Schools Really Failing?

Using Seasonal Comparisons To Evaluate School Effectiveness

Douglas B. Downey\*

Paul T. von Hippel

Melanie Hughes

The Ohio State University

\*Direct all correspondence to Douglas B. Downey, Department of Sociology, 300 Bricker Hall, 190 N. Oval Mall, Columbus, Ohio 43210, (downey.32@osu.edu). Phone (614) 292-1352, Fax (614) 292-6681. This project was funded by grants from the Spencer Foundation, the John Glenn Institute, and the P-12 Project to Downey. We appreciate the comments of Beckett Broh and Brian Powell.

## Abstract

To many it is obvious which schools are failing—those whose students perform poorly on achievement tests. But this method of evaluating schools mixes the effect of school factors (e.g., good teachers) with the effect of non-school factors (e.g., homes and neighborhoods) in unknown ways. As a result, current accountability measures used to determine which schools are “failing”—as mandated by the *No Child Left Behind* legislation—likely underestimate the effectiveness of schools serving disadvantaged populations. We introduce a new measure, “impact,” designed to isolate contributions to learning attributable to schools. Our measure of school *impact* is straightforward – the degree to which schools *increase* their students’ rates of learning when school is in session versus when it is not. With data from the *Early Childhood Longitudinal Study of 1998-98*, we show how conclusions about which schools are failing are altered substantially when we employ measures that isolate school from non-school effects.

Context shapes human behavior. Sometimes state policies emerge, however, that eschew this view. For example, current approaches for evaluating schools draw on market principles from economics but reject the sociological claim that school performance is shaped by multiple institutions. In this essay, we evaluate the current approach used by most states to gauge school effectiveness and show how it is biased against schools serving students from disadvantaged backgrounds. We introduce a new method, school *impact*, which we define as the difference between students' learning rates in and out of school. We suggest that school *impact* provides a fairer measure of school performance.<sup>1</sup>

Accountability systems designed to measure school performance gained favor in most states during the 1990s, a position enhanced even further with the implementation of the federal *No Child Left Behind Act of 2001*. As accountability has become a political reality, the accounting methods employed by states to evaluate school effectiveness merit special attention. Measuring school effectiveness in a compelling way is fundamental to both the effectiveness and legitimacy of current education legislation. One glaring challenge is that most current methods of evaluation assume that schools are the sole influence on children's learning, yet we know this is not right. Even on the first day of kindergarten, schools can be ranked somewhat reliably on their children's test scores, though surely these variations have to do with factors outside of the

---

<sup>1</sup> This review focuses on students' test scores, but a variety of alternative measures of school success have been employed or advocated by researchers. For example, Bliss (1991) discusses holistic outcomes, where effectiveness consists of helping students experience a wide range of content, implying that schools should promote not just rote learning, but a more active problem-solving capacity. Newmann (1991) argues that effectiveness should be defined in terms of the presence of such behaviors and attitudes as self-esteem; racial tolerance; political efficacy; and reduced teen pregnancy, drug abuse and gang participation. Louis and Miles (1991) and Mortimore (1991) argue for multiple indicators of excellence, including information on dropouts, attendance, and student violence. In addition, Rowan (1984) argues that the definitions and measures of school effectiveness vary so substantially that effectiveness should be determined by multiple measures from numerous interest groups and the interrelationships among the different measures should be the subject of further research.

school's control (Lee et. al 2004; Downey, von Hippel, and Broh 2004). And non-school influences continue to matter after children begin formal schooling, both during the academic year and during the summer (Heyns 1978; Entwisle and Alexander 1992, 1994; Downey et. al. 2004). As the Coleman Report highlighted decades ago, families play a dominant role in explaining variations in children's achievement (Coleman et. al 1966).

One might expect that correcting this error would not change things much. After all, schools identified as "failing" under current methods really do appear to be the worst schools. They not only have low test scores but they tend to have high teacher turnover, poor resources, and poor morale (Thenstrom and Thernstrom 2003). Perhaps more importantly, they *look* like failing schools because they often serve the most disadvantaged students. But we will show that if we isolate schools' contribution to students' learning, our ideas about "failing" schools change dramatically. Indeed, by ignoring social context, current methods used to identify "failing" schools are more often wrong than right. In addition, seventeen percent of schools that are currently viewed as satisfactory end up being among the poorest performers when we measure their performance more accurately. We call for a fundamental change in how schools are evaluated. Gauging school performance in a credible way requires a sociologically based approach to measurement.

### THREE MEASURES OF SCHOOL EFFECTIVENESS

We review the current method used for evaluating schools, what we call “achievement,” and then compare it to two alternatives designed to account for students’ varying non-school environments. We make the case for evaluating schools on the basis of *impact*—the degree to which a schools’ students learn faster when they are in school than when they are not.

(1) *Achievement*. At present, *NCLB* allows each state to develop its own proficiency bar, but it provides parameters around some elements of this evaluation. For example, all states are required to test children in math and reading annually between grades 3-8 and at least once between grades 10-12.<sup>2</sup> As one example of how states have responded, the Department of Education in Ohio complies with *NCLB* by using an achievement-bar standard for Ohio schools based on twenty 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* legislation acknowledges varying non-school environments. For example, schools with low test scores are not expected to meet the state’s proficiency bar immediately but can satisfy state requirements by making “adequate yearly progress” for the first several years.<sup>3</sup> In this way, the legislation recognizes that schools serving poor children will need some time to reach the proficiency standards expected of all schools by 2013-2014. But even this accommodation fails to acknowledge fully the importance

---

<sup>2</sup> And in 2007-08 students must be tested in science at least once between grades 3-5, 6-9, and 10-12.

<sup>3</sup> “Adequately yearly progress” definitions vary by state. In Ohio, adequately yearly progress typically means reducing the gap between a school’s or district’s baseline performance (average

of the non-school environment for two reasons. First, the legislation still places the burden of improvement on schools that are below a proficiency bar – a bar constructed without consideration of non-school environments. Second, by 2013-2014 all schools must meet the proficiency bar, regardless of the non-school environments of the children they serve.

The main problem with the achievement standard is that it does not adequately separate school and non-school effects on children's learning. It is likely that a schools' test scores are a function of both school practices (e.g., good teaching and efficient administration) and non-school characteristics (e.g., involved parenting and quality neighborhoods) and so it is not clear 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 and so there is substantial variation in the kinds of students who attend different schools. The challenge, therefore, is to measure the value that schools add *independently from the widely varying non-school factors that influence children's learning.*

Sociologists have documented extensively two characteristics of the home environment: (1) its importance to children's development, and (2) the substantial variation in children's experiences. This variation does not appear to be randomly distributed across schools, given the patterns observed from children just beginning kindergarten in the *Early Childhood Longitudinal Study*. For example, eighteen percent of children entering kindergarten in the U.S. in the fall of 1998 did not know that print reads left to right, where to go when a line of print ends, or where the story ends in a book. At the other end of the spectrum, a small percentage of children beginning kindergarten can already read words in context (West, Denton, and Germino-Hausken 2000). Widely varying skills among children, of course, would not be so problematic for our

---

of 1999-2000, 2000-01, 2001-02 years) and the proficiency bar by 10 percentage points per year between 2003-04 and 2013-14.

goal of measuring school effectiveness if children were randomly distributed across schools, but multi-level models assessing the ECLS data demonstrate substantial between-school variation in children's cognitive skills, even at the very beginning of kindergarten (Downey, von Hippel, and Broh 2004; Lee and Burkam 2002; Reardon 2003). Indeed, in previous work we found that, among children starting kindergarten, 21 percent of the variation in reading test scores and 25 percent of the variation in math test scores was between schools (Downey, von Hippel, and Broh 2004, Table 2, p.622). In short, substantial differences in school achievement 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.

Under achievement methods of evaluating schools, the burden of improvement ends up being disproportionately placed on schools serving children from poor non-school environments, even though it is not yet clear that these schools are doing less well than schools serving children from advantaged environments. Although some schools serving disadvantaged populations may actually be poor-quality schools, without separating school from non-school effects it is difficult to make this evaluation with confidence.

(2) *Learning*. One way to measure school effectiveness that begins to address differences in non-school factors is to gauge how much schools' students improve, rather than where they end up, on an achievement scale. The advantage to this approach is that schools are not rewarded or penalized for the achievement level of their students at the beginning of the year. Under this system, schools serving children with initially high achievement would be challenged to raise performance even further, while schools serving disadvantaged students could be deemed "effective" if its students made substantial progress.<sup>4</sup>

One example of this "learning" approach is the Tennessee Value Added Assessment System (TVAAS), implemented by the state of Tennessee in 1992 to assess its teachers and schools. Under TVAAS, students are measured each year, and data is compiled into a longitudinally merged database linking individual outcomes to teachers, schools, and districts (Chatterji 2002). Estimates of average student achievement progress are calculated for each school and teacher, and the model then determines a school's performance on the basis of estimated gain scores of a school relative to the norm group's gain on a given grade-level test (Kupermintz 2002). North and South Carolina have implemented similar systems (Ladd and

---

<sup>4</sup> One earlier approach involved equalizing schools on observables (e.g., race and socioeconomic status of the student composition), and then predicting student achievement in a regression model. These models would produce expected achievement (the regression line) and could be compared to observed achievement. "Effective" schools were those above the regression line while poor schools were those below the line (Firestone 1991). In this way, achievement expectations were modified relative to the racial and socioeconomic composition of the school. The key limitation to this approach is that observed differences are measured only crudely (e.g., so While "learning" measures have the key advantage of recognizing that schools serve children from different non-school environments, they also force measurement down to the student level so that schools, districts and states track each individual student's progress. An advantage of developing and focusing on student-specific data for accountability is that it schools where mobility is substantial (often urban schools) can be accommodated in some manner. cioeconomic scale) and important differences between schools likely go unmeasured.

Walsh 2002).<sup>5</sup>

Why might yearly gains still be inadequate measures of school effectiveness? In part, it is because children spend the vast majority of their time *outside of school*. Walberg (1984) estimated that the typical 18 year-old American has spent only 13% of his/her waking hours in school.<sup>6</sup> Table 1 presents the proportion of waking hours spent in school estimated for students who miss no days of school across calendar and academic years. During a calendar year, which includes the non-school summer, the proportion is .25. This increases, but only to .32, for an academic year. In short, whether we measure children's gains over a calendar or academic year, the majority of their time is spent outside of school.

Comparing annual test score gains is reasonable, therefore, only if children go home to similar non-school environments between the two measurement periods. But social scientists agree that there is substantial variation in non-school environments and that this variation matters for children's cognitive development. Some children enjoy rich non-school environments with two biological parents (McLanahan and Sandefur 1994), ample educational resources in the home (Teachman 1987), and few siblings with whom to share them (Downey 2001). And these non-school advantages are not just limited to the home. Step outside the door and the advantaged child is exposed to neighbor children who value academic success and neighborhood

---

<sup>5</sup> The TVAAS method is well-suited for identifying years where children make larger (or lower) than expected gains. This information could be used to identify good teachers within a school. But these models are less adept at comparing the learning trajectories of students in disparate schools. If school A has low learning trajectories relative to school B, the TVAAS struggles to identify whether it is due to school or non-school characteristics.

<sup>6</sup>On its surface this estimate appears low, but note that children usually spend only six-seven hours a day in school, and they do not attend school on weekends, holidays, or typically during the summer. Even with perfect attendance, most children spend fewer than one-half the days a year in school (180/365), and children's pre-kindergarten years are spent primarily in "non-school" environments.

adults who both enforce norms regarding school-related behaviors (e.g., finish your homework before playing) and, via their professional jobs, serve as concrete evidence that schooling efforts payoff (Brooks-Gunne, Duncan, and Aber 1997; Wilson 1996). Even during the school day, the advantages (or disadvantages) of a non-school environment can matter. Parents vary widely in the extent to which they are involved in helping their children with homework, meeting with teachers, and actively promoting their child's interests at school (Lareau 2000).

As one example of how much home environments vary in cognitive stimulation, Hart and Risley (1995) observed that among children six months to three years old, welfare children had 616 words per hour directed to them compared to 1,251 for working-class and 2,153 for professional-family children. The authors extrapolated these patterns to highlight the cumulative impact of this disparity up until age four. Inferring slightly further, until the beginning of kindergarten, the results from this study suggest that the average child in a professional family will have had 61 million words directed their way before beginning kindergarten, compared to 36 million for working-class and only 18 million for welfare children.<sup>7</sup> Given such varying exposure to language, the large skill gaps among children beginning kindergarten are not surprising.

While measuring school effectiveness as *learning* represents a substantial improvement over the *achievement* approach, the main limitation of *learning* is the assumption that non-school influences are similar for all children during the period when learning is measured—typically a calendar year. In fact, however, children spend most of their time outside of school (about three-quarters of their waking hours during a calendar year) and so, through no effort of their own,

---

<sup>7</sup> These figures assume that children are awake for 14 hours a day and that observations of parent-child interaction made during six months to three years remain constant until the child enters kindergarten at 5 ½ years of age.

schools serving children with advantaged non-school environments will more easily register *learning* gains than will schools serving children with poor non-school environments.

Most value-added measures of school performance attempt to measure some of the characteristics of students' non-school environments and statistically adjust their expectations of schools accordingly. But to isolate school effects successfully with this approach, all relevant differences in non-school environments would have to be identified and measured perfectly.<sup>8</sup> Because that is a nearly impossible task, Meyer concludes that “[t]he principal obstacle to developing a high-quality indicator system is the difficulty in collecting extensive information on student and family characteristics” (Myer 1996:210). Our own previous work has highlighted the limitations of strategies of equalizing children’s non-school environments by adjusting for covariates. Typical measures of the non-school environment (e.g., parents’ socioeconomic status, family structure, race, gender) explain only thirty percent of the variation in children’s cognitive skills when they begin kindergarten and only *one percent* of the variation in summer growth rates (Downey, von Hippel, and Broh 2004). In short, models employing typical covariates explain so little of the variation in non-school learning rates that adjusting children’s non-school environments in this way seems implausible. Our alternative strategy for isolating school from non-school effects circumvents this problem.

---

<sup>8</sup> Rubenstein, Stiefel, Schwartz, and Amor (2004) refer to this approach as Adjusted Performance Measures (APMs). Using this approach, a school’s test scores could be regressed on a set of independent variables thought to be out of the school’s control (e.g., percentage of students eligible for free lunch). A school’s APM is then its residual value or the gap between the expected performance predicted by the regression equation and the observed performance. But as the authors point out, “APMs implicitly assume that all of the estimated error reflects relative school efficiency or inefficiency. To the extent that the error term captures other factors, such as measurement error or the effects of unobserved or omitted variables, the residuals may under- or overestimate school efficiency” (Rubenstein et. al. 2003, p.59).

**(3) Impact.** Our method for evaluating school effectiveness—what we call “impact” — represents an attempt to isolate school effects more completely by considering how students’ learning rates change when they are in school versus when they are not. The logic behind the impact measure is straightforward. Our reasoning is that, whereas school-year learning is a product of both non-school and school factors, summer learning is the product of non-school factors exclusively. The difference between summer and school-year learning rates, therefore, represents the school’s impact, or effect on learning. A high-impact school is one that boosts children’s learning rates substantially above the rates observed when the children were not in school. By focusing on the degree to which schools increase the rate at which their students learn when not in school, we hope to more successfully separate school from non-school effects on learning.

A key advantage of this approach is that we circumvent the formidable task of trying to measure and statistically adjust for all of the different aspects of children’s non-school environments. While past research has tried to account for non-school differences using measures of poverty, ethnicity, and family structure among other things (Clotfelter and Ladd 1996; Ladd and Walsh 2002), it is rarely clear whether a sufficient number of non-school confounders have been measured and measured well. By focusing instead on summer learning, our measure attempts to capture what we need to know about children’s learning opportunities outside of school without the challenge of identifying all of the many features of the non-school environment that matter. Another advantage of our approach is that we are not forced to make the assumption that variations in learning are solely explained by *environmental* conditions. Even non-environmental effects on learning (e.g., variations in innate motivation level) are better

accounted for when we make summer/school year comparisons.<sup>9</sup>

An estimate of impact requires seasonal data—data collected at both the beginning and end of successive years of school. Seasonal data are quite rare in educational research, but typically quite revealing. For example, previous researchers have noted that gaps in cognitive skills across socioeconomic status grow primarily during the summer, when children are out of school, and are likely reduced during the school year (Heyns 1978; Entwisle and Alexander 1992, 1994; Downey, von Hippel, and Broh 2004; Reardon 2003). The notable advantage of seasonal data is that it allows for an estimate of children’s rate of cognitive growth during the summer, when children are not in school. Knowing how fast children learn when exposed to their non-school environment provides critical leverage for isolating school effects.

The impact measure requires assumptions. First, the measure assumes that there is little contamination between seasons—mainly that school characteristics do not have important influences on subsequent summer learning. If summer learning depends in an important way on school practices from the previous year, then summer learning fails to provide an uncontaminated measure of the non-school environment. Because there is so little seasonal data available, this is a difficult assumption to evaluate with confidence, but the current empirical information suggests that the assumption is reasonable. Using the *ECLS-K* Georgies (2003) reported no relationship between kindergarten teacher practices or kindergarten classroom characteristics and children’s summer learning. And in our own supplemental analyses of *ECLS-K*, we found that socioeconomically advantaged children were more likely to receive a summer booklist from their kindergarten school, but *less* likely to receive a preparation “package” for first grade than their disadvantaged counterparts. Importantly, neither of these school practices

---

<sup>9</sup> Note that “learning” measures fail to account for the possibility that schools may serve

were related to cognitive gains during the summer. Overall, the “wall” between school and non-school effects may be porous, but current information suggests that the distinctions between seasons are arguably strong enough that seasonally-based estimates represent an important step toward isolating school from non-school influences on learning.

Second, the *impact* measure assumes that the non-school rate of learning observed in the summer continues at a constant rate through the school year. There are reasons to suspect that non-school effects may weaken during the school year relative to the summer for the obvious reason that children spend less time exposed to the non-school environment during the school year. For this reason we will test the robustness of our school rankings via impact by adjusting the fraction of the non-school rate of learning that we subtract from the school-year rate. A more subtle and potentially troublesome possibility is that the non-school effect on learning varies across seasons *and* across comparison groups. Suppose high-SES parents, for example, invest substantially in the summer but then relatively less so during the school year while low-SES parents produced the opposite seasonal pattern.<sup>10</sup> This kind of scenario would produce predictable biases in the *impact* measure, underestimating school impact for schools serving high-SES families and overestimating the performance of schools serving low-SES parents. While strong differences in the ratio of non-school inputs across seasons and across comparison groups may seem unlikely, little is known about this possible source of bias. While all evaluative systems require some assumptions, we note that the assumption required for the impact measure is more modest than the assumptions needed for achievement and learning

---

children with varying learning trajectories.

<sup>10</sup> Most of what we know about parental involvement in children’s schooling suggests that this pattern is unlikely. Socioeconomically advantaged parents maintain active involvement in their children’s lives during the academic year by helping with homework, volunteering in classes, and attending school activities and parent-teacher conferences (Lareua 2000).

models (i.e., that non-school factors are irrelevant to achievement and learning).

## METHODS AND RESULTS

*Data.* We use the *Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K)*, a survey administered by the National Center for Education Statistics, U.S. Department of Education (National Center for Education Statistics 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.

992 schools were visited in the fall of kindergarten (time 1), the spring of kindergarten (time 2) and the spring of first grade (time 4). Among these 992 schools, 309 were randomly selected for an extra visit in the fall of first grade (time 3). We focus on the 309 schools with time 3 data, since these are the schools for which we can estimate first grade and summer learning rates. On average, 19 children were tested in each school (the median was 20), but in individual schools as few as one or as many as 25 students were tested.

We present results based on evaluating schools on the basis of their reading and math test scores. The reading assessment tests 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 is 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.

Like many surveys, *ECLS-K* is missing a number of values. For the most part, values are missing because of nonresponse or unavailability, but we also deleted a small fraction of values because of obvious recording errors, or because a child had transferred from his or her original school. We compensated for missing values using *multiple imputation with deletion*. In *multiple imputation* (Rubin 1987), each missing value is replaced with several (in our case, six) plausible imputations. *Multiple imputation with deletion* (von Hippel 2004a) is a refinement that improves statistical efficiency by deleting imputations of the dependent variable (in our case, test scores).<sup>11</sup>

**Analytic Strategy.** We begin by ranking schools in terms of achievement and creating an arbitrary cutoff point of “failing” schools as those scoring among the bottom twenty percent. Of course it is well known that states have created widely varying criteria for determining “failing” status, and so in some states the percentage of “failing” schools is substantially lower than twenty percent, while in others it is greater. Our cutoff point for this national sample of schools

---

<sup>11</sup> Independent variables with missing values included student body characteristics (percent minority, percent free lunch, percent reduced lunch), dates for the beginning and end of kindergarten and first grade, and test dates. (School dates and test dates were needed to calculate how long each child had been exposed to KINDERGARTEN, SUMMER, and FIRST GRADE at the time of each test.) These are all school-level variables; test dates have 91% of their variance between schools, and the other variables vary entirely between schools. Before imputation, we constructed a school-level data set which contained all of the school-level variables as well as child-level variables averaged up to the school level. In this school-level data set, missing values were imputed under a multivariate normal model.

is not meant to map directly onto each state’s definition. Rather, we define the bottom twenty percent of schools as “failing” here as a heuristic tool to highlight how the different methods of evaluating schools produce widely varying conclusions.<sup>12</sup>

We then calculate each schools’ *learning* and *impact* scores and present new rankings of the schools according to these different methods. Our interest is in how much things change. For example, what percentage of “failing” schools, as defined by *achievement*, are no longer “failing” once we move to *learning* and *impact* definitions? In addition, what percentage of schools viewed as successful under the *achievement* model, would be considered “failing” under *learning* or *impact* definitions?

### ***Model specification***

We estimate cognitive growth by fitting a *multilevel model for change* (Singer and Willett 2003; Raudenbush and Bryk (2002)). In our 3-level model, we view test scores (level 1) as nested within children, and children (level 2) as nested within schools (level 3).

At level 1, we model each test score  $Y$  as a linear function of the months that child  $c$  in school  $s$  has spent in KINDERGARTEN, SUMMER, and FIRST GRADE at the time of test  $t$ :

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

The slopes  $\alpha_{1cs}$ ,  $\alpha_{2cs}$ , and  $\alpha_{3cs}$  are monthly rates of learning during kindergarten, summer, and first grade. The meaning of the intercept  $\alpha_{0cs}$  depends on the coding of the variables KINDERGARTEN, SUMMER, and FIRST GRADE. In the most obvious coding, KINDERGARTEN would begin at zero and increase to 9.5 or so by the end of the kindergarten year. In our coding,

---

<sup>12</sup> In supplemental analyses we explored alternative cutoff points but our substantive conclusions

however, KINDERGARTEN begins at  $-9.5$  or so and increases to zero at the end of the kindergarten year. The other variables are coded in a similar way. The intercept  $\alpha_{0cs}$ , then, represents the child's score on the last day of first grade, when  $\text{KINDERGARTEN}=\text{SUMMER}=\text{FIRST GRADE}=0$ . (This last-day score is an extrapolation; it is not the same as the final test score, because the final test was typically taken one to three months before the end of first grade.)

The residual term  $e_{ics}$  is measurement error—the difference between the test score  $Y$  and the child's true achievement level. The variance of the measurement error can be derived from test-reliability estimates in Rock and Pollack (Rock and Pollack 2002); Table 1 gives the requisite calculations. The measurement error variance changes little from one occasion to the next; we obtain similar results whether we assume heteroscedasticity or not.<sup>13</sup>

At level 2 of the model, we break each child's parameters  $\alpha_{cs}$  into school- and child-level components  $\beta_s$  and  $a_c$ :

$$\alpha_{0cs} = \beta_{0s} + a_{0c}$$

$$\alpha_{1cs} = \beta_{1s} + a_{1c}$$

$$\alpha_{2cs} = \beta_{2s} + a_{2c}$$

$$\alpha_{3cs} = \beta_{3s} + a_{3c}$$

In each equation, the first component  $\beta_s$  represents the average parameters for school  $s$ , and the

---

were the same.

<sup>13</sup> Walsh and Ladd (2002) noted that value-added approaches for measuring school effectiveness in South Carolina and North Carolina suffer from a lack of attention to measurement error. Using an instrumental variable approach for correcting for measurement error, they estimate that about two-fifths of the correlation between school effectiveness and SES or race is attributable to measurement error. They conclude that “[w]ithout the correction, schools serving low-performing students and students from disadvantaged backgrounds are viewed as being less

second component  $a_c$  represents the departure of child  $c$  from the school  $s$  average. The child-level effects ( $a_{0c}, a_{1c}, a_{2c}, a_{3c}$ ) are assumed to be random with a multivariate normal distribution.

At level 3, we break each school's parameters  $\beta_s$  into fixed and school-level components  $\gamma$  and  $b_s$ :

$$\beta_{0s} = \gamma_0 + b_{0s}$$

$$\beta_{1s} = \gamma_1 + b_{1s}$$

$$\beta_{2s} = \gamma_2 + b_{2s}$$

$$\beta_{3s} = \gamma_3 + b_{3s}$$

In each equation, the first component  $\gamma$  represents the grand average of the parameter values, and the second component  $b_s$  represents the departure of school  $s$  from the grand average. The school-level effects ( $b_{0s}, b_{1s}, b_{2s}, b_{3s}$ ) are assumed to be random with a multivariate normal distribution.

### ***Parameter estimation***

For school  $s$ , the average *achievement* level at the end of first grade is  $\beta_{0s}$ , the average first grade *learning* rate is  $\beta_{3s}$ , and the average *impact* is the difference  $\beta_{3s} - \beta_{2s}$  between the first grade and summer learning rates. We estimated these quantities using empirical Bayes methods (also known as BLUPs), which compensate for measurement error and sampling error by shrinking extreme values toward the grand mean (Raudenbush and Bryk 2002b; Robinson 1991). Empirical Bayes methods ensure that a school with little information—i.e., a school where few children were tested—is unlikely to have extreme parameter estimates (von Hippel 2004b).

---

effective than they really are and those serving high-performing students are viewed as being

### *Evaluations of School Effectiveness Across Three Indicators*

Using these estimates, we identified the schools that were in the bottom quintile (“failing”) with respect to achievement, learning, and impact. Results for reading are shown in Table 2 and graphically in Figure 1., highlighting the disagreement among these criteria.<sup>14</sup> Among the schools that were in the bottom quintile with respect to reading *achievement*, just 52 percent (32 of 61) were in the bottom quintile with respect to reading *learning*, and only 30 percent (18 of 61) were in the bottom quintile with respect to reading *impact*. Figure 1 shows how some schools that were failing with respect to achievement were actually above average with respect to learning or impact. Correlations between reading achievement and learning (.51) and reading achievement and impact (.12) are smaller than the correlation between reading learning and impact (.71).

If we evaluate schools on the basis of math skills, the disjuncture between achievement, learning, and impact measures is even greater (Table 3 and Figure 2). Among the schools that were in the bottom quintile with respect to math *achievement*, just X percent ( of ) were in the bottom quintile with respect to math *learning*, and only Y percent ( of ) were in the bottom quintile with respect to math *impact*. Figure 2 demonstrates how schools vulnerable to the “failing” label in terms of math achievement often look much better when we use an indicator (learning or impact) that more persuasively isolates schooling’s contribution to learning.

---

more effective than they really are” (Walsh and Ladd 2002:12).

<sup>14</sup> Initially we feared that the correlation between the criteria might be attenuated by the need to use estimated rather than true school-level values for achievement, learning, and impact. In fact, however, the correlation between the empirical Bayes estimates of achievement, learning, and impact are very close to the estimated correlations between the true school-level values. This may be a property of empirical Bayes estimators—consistent estimation of variances and covariances? CHECK.

Correlations among math achievement and learning ( . ) and math achievement and impact ( ) are compared to the correlation between math learning and impact ( ).

### *What Kinds of Schools Tend to Fail?*

The kinds of schools that look poor on our measures change depending on the criterion for failure (Table 3). If we use achievement as a criterion, the characteristics of failing schools are quite familiar. Schools in the bottom quintile tend to be urban and public, with high populations of minority children and children eligible for free or reduced lunch. If we use *learning* as a criterion, however, the characteristics of failing schools are less clear. Public schools are actually less vulnerable to failure than Catholic schools, and urban schools are only a little more vulnerable than suburban schools. Failure is still associated with high-minority and high-poverty populations, but the association is weaker. Finally, if we use *impact* as a criterion, the results are similar to those for learning, yet school performance is even less related to background characteristics.<sup>15</sup> Note, however, that our impact measure is positively correlated with the achievement measure. Poor schools, by the achievement measure, *tend* to be poor schools by the impact measure too. And so while the impact measure changes our way of thinking about which schools are “failing,” it has significant overlap with the achievement approach.

### *Validity and Reliability*

Children have good days and bad days and so their performance on a math or reading test at one point in time does not typically indicate their skills perfectly. This measurement error

causes test scores to wobble around their true values, a problem that may appear modest if we believe that children with good days tend to cancel out the children with bad days. But this only works when we have a large enough number of students from a school so that we can be confident that the “canceling out” is really working. Kane and Staiger (2002) showed how schools with small populations are especially vulnerable to being misclassified, solely because the reliability of estimates is weak in small schools.

While we contend that evaluations based on *learning* (two measurement points) or *impact* (three measurement points) have greater validity than *achievement* scores (one measurement point), we recognize that they may have less reliability. Each time we gauge children’s skills we do so with error and so measures of school estimates that depend on more assessment points (i.e., *learning* and *impact*) are vulnerable to greater measurement error than estimates based on fewer assessment points (achievement). It is worth asking, therefore, if the loss in reliability is worth the gain in validity? This problem is especially relevant to analyses of the *ECLS-K* data because we were only able to assess the test scores for a sample of children in each school (~ 20). Would the substantial misclassification that we report above be reduced if we could assess every child in each school? And if accountability systems switched from an achievement to learning or impact, would the increase in validity be worth the drop in reliability?

[expand this section]

## DISCUSSION

A simple common-sense observation—that children are influenced in important ways by

---

<sup>15</sup> The one difference is that high-impact schools actually have a higher percentage of free-lunch

their non-school environments—undermines the achievement method for evaluating schools. Current methods systematically underestimate the quality of schools serving disadvantaged students and overestimate the quality of schools serving the advantaged because they ignore the overwhelming amount of time that children spend outside of school. 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. Confidently identifying “failing” schools, therefore, requires a method of evaluation can separate school from non-school effects on learning.

Our contribution has been to highlight conceptual problems with *achievement* and *learning* methods of evaluating schools, and to show how our ideas about which schools are failing change substantially when we more successfully recognize non-school factors. Rigorously isolating school from non-school effects is crucial to producing unbiased estimates of school effectiveness. The striking pattern that emerges from our results is that schools serving socioeconomically advantaged children are identified as attending much better schools under the achievement model, modestly better schools by the learning model, and comparable schools by the impact measure. This pattern of results suggests that, as we account for greater proportions of non-school factors, the schools serving advantaged students look more and more like other schools.

Based on these results it is likely that many schools are mislabeled as “failing” by current state standards. In these mislabeled schools, teachers are improving their students’ rates of learning (above that observed when the children are not in school) as much, and sometimes more, than teachers serving advantaged students. The extent of the bias is substantial—our analyses of reading suggest that 70 percent of currently labeled “failing” schools are not really

---

students than do low-impact schools.

failing (as defined by our 20 percent cutoff point). Many teachers and administrators working in schools serving disadvantaged children face a variety of challenges including scarce resources, large classes, and little parental involvement. Despite these conditions, a surprising number of professionals serving disadvantaged students appear to be doing a good job, much better than previously thought.

The validity of school measures is critical to the success of accountability systems because making this information publicly available is supposed to pressure school personnel to improve. But our results suggest that the information currently available to construct the notion of “good school” is substantially flawed. Poor information reduces market efficiency in the school case by too often sending parents away from solid schools serving children from disadvantaged backgrounds and insufficiently pressuring schools serving children from advantaged backgrounds. The error in achievement measures is so great that accountability systems based on achievement may result in a substantial misdirection of resources.

In practice, holding schools accountable via an *impact* measure raises several issues. First, to produce the necessary seasonal estimates, children would need to be tested twice a year. Currently, *NCLB* requires once a year testing between grades 3-8, so a seasonal method would double that requirement during the early years, an unattractive option for most policymakers, school personnel and parents. A practical alternative would be to maintain the same number of assessments but alter their timing. For example, policymakers could decide that it is more important to have quality information about school quality at two points in time, rather than poor information about school quality six times.<sup>16</sup>

---

<sup>16</sup> two sets of seasonal data (assessments at the end of third grade, the beginning and end of fourth, the end of seventh, and the beginning and end of eighth) would require the same number

We have argued here that achievement measures have important limitations when the goal is to hold schools accountable. But for other reasons it may still be useful to maintain publicly available achievement information. For example, if our interest shifts from wanting to evaluate schools fairly to wanting to know if children are reaching a particular level of proficiency, then achievement information could be useful for directing additional resources toward poor achieving schools. High impact schools with low achievement might be especially attractive schools in which to invest additional resources, given that they appear to be operating efficiently.<sup>17</sup>

While our focus has been on isolating school from non-school influences on school performance, it is important to note that “school” effects in our models may not be appropriate for accountability measures. The school effects, as indicated by “impact,” are likely a function of: (1) school characteristics partly under the control of school personnel (e.g., teachers and administrators working harder and smarter), and (2) school characteristics partly or mostly beyond the control of school personnel (e.g., academically strong peers, attracting and retaining high-quality teachers, and school resources). A fair measurement tool for holding school personnel responsible for student performance would require further removing the effect of school characteristics for which the school has little control. Of course, the degree of control over some of these school characteristics is difficult to gauge. For example, a school may partly influence the kinds of teachers it attracts and retains, and it may shape its resources at some level

---

of assessments as six annual tests, but provide more accurate information about the school’s effectiveness.

<sup>17</sup> Would the public accept a value-added measure of schools rather than the more straightforward achievement measure? Clearly the estimation procedure is complex and requires advanced statistical tools. Of course, those states employing other kinds of value-added measures of school performance have trained statistical experts to perform the analyses with

through fundraising and entrepreneurial initiatives. And one could even make the case that, once market-based reforms have been in place long enough, schools influence the kinds of students that they attract. School personnel's control over these factors may vary, but it is never complete. The direction of this bias is again predictable—schools serving the disadvantaged are likely evaluated too harshly by our impact model.

Our study illustrates that by simply focusing on one problem, the need to distinguish school from non-school effects, we uncover substantial bias against schools serving the disadvantaged. This may be the largest and most overlooked source of bias in current estimates of school effectiveness. Further isolating that component of the school effect that is arguably under school personnel's control is beyond the scope of this study, but remains an important next step before we can confidently develop a measure that would fairly evaluate school personnel. Measuring school effectiveness as achievement may not only be biased against schools serving the disadvantaged, it may undermine a key goal of the *NCLB* legislation—to reduce racial/ethnic and socioeconomic performance gaps. If schools serving the disadvantaged are evaluated on a biased scale, their teachers and administrators are likely to respond like workers in other industries when they are evaluated unfairly, with frustration and reduced effort (Hodson 2001). Our call is for no special favors for these schools but rather that they be evaluated on the same standard as other schools. One way to consider whether an evaluative tool is biased is to consider the question, “Do teachers and administrators feel as if they have an equal chance of success regardless of the school in which they happen to teach?” Right now the answer to that question is “no.” Teachers of advantaged children perceive little chance of failure while the teachers of disadvantaged children perceive little chance of success, contributing to the incentive

---

little public outcry. And we would point out that the impact measure has a conceptually appealing simplicity, the difference between rates of learning in versus out of school.

teachers have to leave challenging schools. Under a fair system, a school's chances of success should not depend so heavily on the kind of students it serves.<sup>18</sup>

Future work must also grapple with a potential source of bias that works in the opposite direction, favoring schools serving the disadvantaged. Students with high scores at the beginning of the school year may have less room to improve than students with low scores. If this is the case, then schools serving advantaged children may have more difficulty scoring well on an *impact* measure (or, for that matter, on a learning measure) than do schools serving disadvantaged children. As noted earlier, for the ECLS-K data, the *National Center for Education Statistics* attempted to reduce the problem of ceiling effects by routing children toward appropriate-level tests of varying difficulty. While it appears that this strategy was successful (no clustering near the top scores) it is difficult to know if it successfully allowed students to *gain* equal amounts on the scale during the academic year, a challenge that may affect the reading scale more than the math scale (Condrón, 2005).<sup>19</sup> In addition, initial clustering near the *bottom* of the reading scales is unavoidable since, as noted earlier, about a fifth of children

---

<sup>18</sup> And while we may occasionally rejoice in the remarkable increases in achievement test scores of some schools serving disadvantaged children, these few exceptions do not mean that outside of school factors are inconsequential. As Rothstein explains: "Those who insist that poverty does not cause low achievement usually say that failure is caused instead by low standards and expectations, or inadequate testing and accountability. Yet we all know schools with high standards where some children fail, and we all know inadequate schools from which some students nonetheless emerge successful. These exceptions do not make it fashionable to conclude that standards and accountability must not matter. We recognize that, on average, students will do better in schools with high standards, even though not every student will do so." (Rothstein 2004: p.62-63).

<sup>19</sup> Condrón (2005) shows that ECLS-K students scoring high at the beginning of first grade tended to register lower gains than students starting with low scores, indicative of a ceiling effect. Perhaps because of a broader scale, however, initial math scales are not related to math gains.

simply have no skills related to reading.<sup>20</sup>

It is also possible that a developmental ceiling prohibits advantaged children from demonstrating greater “impact” in our models. High-performing six-year olds may be able to learn only so much more reading, regardless of school quality. *Impact* could appear relatively modest for schools serving advantaged children, therefore, because advantaged children more frequently bump up against this developmental ceiling. This is merely speculation at this point, however, and a plausible alternative is that schools serving advantaged children are not challenging their students to make the next leap.

Is our new method of evaluation just a thinly veiled attempt to avoid holding schools accountable? Absolutely not. We see value in accountability, but only if schools are held accountable for what they can reasonably control. Accountability is undermined if the measure of school effectiveness is unreasonable. And note that schools serving advantaged students have something to gain from our new way of evaluating schools. If parents of economically advantaged children want their children to receive a good education, then they should be challenging their school to teach their child more than the child would learn if not in school. In our models, seventeen percent of schools with high reading achievement test scores (top four quintiles) scored poorly on our “impact” measure (bottom quintile). The children in these schools should be challenged more so that they avoid the “soft bigotry of low expectations.”

These results bring up several issues regarding measuring school effectiveness. First, while it is not surprising that our “impact” measure produces different results than the “achievement” approach, one might still expect schools serving advantaged children to produce

---

<sup>20</sup> This floor problem also presents difficulties, especially when measuring children’s initial gains in kindergarten. The ECLS-K tests likely did not allow for some students to score as low as their

greater “impact” than those serving disadvantaged children. After all, schools serving advantaged children often enjoy more experienced teachers, smaller classrooms, newer textbooks, and other advantages in resources that we think influence learning. The fact that advantaged children in our models enjoy no more “impact” from schooling than disadvantaged children suggests that either these resource advantages matter less than previously thought, or that advantaged children do not benefit more from schooling for other reasons.<sup>21</sup>

While “impact” is arguably a fairer measure of schools than current approaches, important conceptual and practical issues remain. For example, our models assume that the nonschool impact is the same when school is in session and when it is not—i.e., when the school “faucet” turns on, the nonschool “faucet” continues flowing at a constant rate (Entwisle and Alexander 1997). This assumption is debatable. The simplest counterargument is to point out that children spend less time in their nonschool environments during the school year than during the summer. On these grounds, it can be argued that only a *fraction* of the summer learning rate should be subtracted from the kindergarten or first grade learning rates. Although we are

---

true score, thereby underestimating their gains. This source of bias would work against schools serving the disadvantaged.

<sup>21</sup> One possibility is that, while the ECLS-K test of reading skill may not be limited by a methodological ceiling, teachers produce an informal one. To the extent that teachers are primarily concerned with raising *all* of their students’ skills to the level needed for grade-level advancement, they may disproportionately direct their energies toward low achievers. It is also possible that a developmental ceiling prohibits advantaged children from demonstrating greater “impact” in our models. High-performing six-year olds may be able to learn only so much reading, regardless of school quality. Impact appears relatively modest for schools serving advantaged children, therefore, because they more frequently bump up against this developmental ceiling. If this were the case, school A serving advantaged children may exhibit less “impact” than school B serving disadvantaged students, despite similar teaching quality. An equally plausible alternative, however, is that schools serving advantaged children are not challenging their students as much as previously thought. Of course, this developmental ceiling is speculative. If high expectations are critical to student performance, then we should be careful not to assume that high-performing children cannot learn more than they currently are. At any

exploring this possibility, the appropriate fraction is difficult to estimate since it requires that we know the seasonal distribution of nonschool inputs. For example, we know little about whether parents invest more energy in their children's learning during the summer versus school year.

It should be noted, however, that although the assumption of our current models--that non-school inputs are constant across seasons—may eventually be shown to require modification, this issue is likely of modest importance for isolating school effects. For our current models to be biased, the proportion of non-school inputs would need to change across seasons *and* vary across the families attending advantaged and disadvantaged schools. For example, “impact” estimates would be biased if advantaged parents made extraordinary inputs into their children's lives during the summer, but then backed off considerably during the school year, while disadvantaged parents maintained a relatively constant level of low inputs. Under these conditions, “impact” would underestimate the contributions of advantaged schools.

Our motivation for introducing the impact measure of school effectiveness was to more persuasively isolate school from non-school influences on children's learning. But we warn that our impact measure, while more persuasively isolating school effects, may not represent school effects that are readily under the control of school personnel. For example, if students' learning depends heavily on teacher quality, and teachers consistently prefer schools serving socioeconomically advantaged rather than disadvantaged children, then school impact may be shaped, at least in part, by factors not under a principal's control. Similarly, some have argued that children learn more when they are surrounded by academically skilled peers, another

---

rate, if ceiling effects do shape the patterns we report here, this bias is surely modest compared to that produced by ignoring non-school factors.

characteristic of the school experience over which school personnel have little control.<sup>22</sup> While we have provided a tool for isolating school from non-school effects on student learning, to persuasively gauge the effectiveness of school personnel for accountability reasons, researchers would need to isolate further those aspects of school impact that are arguably within the control of teachers and administrators.

This study and previous work done with seasonally collected data represent a challenge to the claim that important social problems (e.g., inequality) are largely a function of the inadequacy of schools. While acknowledging that some schools really are performing poorly,<sup>23</sup> several patterns point us away from schools as the engine of inequality. First, as noted earlier, substantial inequality in children's reading and math skills exists at the very beginning of kindergarten, before schools can have a chance to matter (Lee and Burkam 2002). Variations in cognitive skills among young children, therefore, largely *reflect* the substantial inequality in children's non-school environments. Second, seasonally collected data consistently demonstrates that inequality in math and reading skills grows faster in the summer, when school is out, than it does during the 9-month academic year (Heyns 1978; Entwisle and Alexander 1992, 1994; Downey et. al., 2004). While schools may not provide equal opportunities for children, they are markedly more equal than non-school environments. The result is that schools are more part of the solution than the problem of inequality.

Finally, the results reported here underline the challenge of reducing inequality via

---

<sup>22</sup> Of course, school personnel have some control over the academic skills of students in their school. After all, they teach them. Our point is that the bulk of variation between schools in peer's skills is out of their control, especially during the early grades (Downey, von Hippel, and Broh, 2004).

improving schools. Closing the gap in school performance between poor and middle-class children would require more than just providing similar school opportunities—they largely have that already. Little in our data, for example, suggests that switching the teaching and administrative personnel from schools with high achievement test scores to those with low achievement test scores would do much to improve disadvantaged children’s experiences. To reduce inequality via schooling, where children spend only a fraction of their time, would require that disadvantaged children enjoy *substantially* better school experiences than advantaged children. Schools can only do so much to compensate for American children’s widely varying non-school environments. What is the role of school reform in society? Is it the opiate of the social reformer with school fads ebbing and flowing with little consequence for social problems? Is it the conscious and pernicious attempt by those in power to divert attention away from exploitive economic arrangements? Our paper is not about the motives of those groups that favor or oppose current accountability measures. But we note that the current system is based on several ideas about how the world works that require further thought. Is social inequality primarily driven by the fact that poor children have poor school opportunities? Downey, von Hippel, and Broh (2004) maintain that, overall, schools as an equalizing force. Large gaps in cognitive skills are evident at the beginning of kindergarten and these grow fastest when school is not in session. If anything, schools appear to reduce inequality.

---

<sup>23</sup> Footnote describing the percentage of schools doing poorly on all three measures for both reading and math (these have to be really lousy schools) and a brief description of these schools in terms of percent free lunch, percent minority, and public/private.

## REFERENCES

- Allison, Paul D. 2002. *Missing Data*. Thousand Oaks, CA: Sage.
- Black, Sandra (1999). "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114(2): 577-599.
- Bliss, J. R. 1991. Pp. 43-57 in *Rethinking Effective Schools: Research and Practice*. Bliss, J. R., W. A. Firestone, C. E. Richards, Eds. Englewood Cliffs, NJ : Prentice Hall.
- Booher-Jennings, Jennifer Lee. 2004. "Responding to the Texas Accountability System: The Erosion of Relational Trust." Paper presented at the Annual Meetings of the American Sociological Association, San Francisco.
- Brooks-Gunn, Jeanne, Greg J. Duncan, and J. Lawrence Aber. 1997. *Neighborhood Poverty: Context and Consequences for Children*. New York: Russell Sage Foundation.
- 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(3): 345-86.
- Denton, Kristin and Jerry West. 2002. Children's Reading and Mathematics Achievement in Kindergarten and First Grade, NCES 2002-125. Washington DC: U.S. Department of Education, National Center for Education Statistics.
- Downey, Douglas B. 1995. "When Bigger is Not Better: Family Size, Parental Resources, and Children's Educational Performance." *American Sociological Review* 60:747-761.
- Downey, Douglas B., Paul T. von Hippel, and Beckett Broh. 2002. "Are Schools the Great Equalizer? Using Seasonal Comparisons to Assess Schooling's Role in Inequality." Paper presented at the American Sociological Association Meetings in Chicago, IL.
- 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.
- Entwisle, Doris R. and Karl L. Alexander. 1994. "The gender gap in math: Its possible origins in neighborhood effects." *American Sociological Review* 59:822-838.
- Georgies, Annie. 2003. "Explaining Divergence in Rates of Learning and Forgetting among First Graders." Paper presented at the American Sociological Association Meetings in Atlanta.
- Hart, Betty and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experiences of Young American Children*. The University of Kansas: Paul H. Brookes Publishing Co.
- Heyns, Barbara. 1978. *Summer learning and the effects of schooling*. New York: Academic Press.
- 1987. "Schooling and cognitive development: Is there a season for learning?" *Child Development* 58:1151-1160.
- Hodson, Randy. 2001. *Dignity at Work*. New York: Cambridge University Press.
- Hu, D. 2000. "The Relationship of School Spending and Student Achievement When Achievement is Measured by Value-Added Scores." Ph.D. dissertation. Nashville, TN: Vanderbilt University.
- Kupermintz, H. 2002. "Value-Added Assessment of Teachers: The Empirical Evidence." Pp. 217-234 in *School Reform Proposals: The Research Evidence*. Alex Molnar, Ed. Greenwich, CT: Information Age Publishing.
- 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.
- Lee, Valerie E. and David T. Burkam. 2002. *Inequality at the Starting Gate: Social Background*

- Differences in Achievement as Children Begin School*. Economic Policy Institute: Washington, DC.
- Lareau, Annette. 2000. *Home Advantage: Social Class and Parental Intervention in Elementary Education*. Oxford: Rowman and Littlefield.
- Little, Roderick J. A. 1992. "Regression With Missing X's: A Review." *Journal of the American Statistical Association* 87(420):1227-37.
- Louis, K. S. and M. B. Miles. 1991. "Managing Reform: Lessons From Urban High Schools." *School Effectiveness and School Improvement* 2(1):75-96.
- McLanahan, Sara, and Gary Sandefur. 1994. *Growing Up with a Single Parent: What Hurts, What Helps?*
- Meng, X. L. "Multiple Imputation Inferences With Uncongenial Sources of Input." *Statistical Science* 10:538-73.
- Mortimore, P. 1991. "Effective Schools From a British Perspective: Research and Practice. Pp. 76-90 in *Rethinking effective schools : research and practice*. Bliss, J. R., W. A. Firestone, C. E. Richards, Eds. Englewood Cliffs, NJ: Prentice.
- Meyer, Robert H. 1996. "Value-Added Indicators of School Performance." Pp. 197-223 in *Improving America's Schools: The Role of Incentives* (Eds. Eric A. Hanushek and Dale W. Jorgenson). National Academy Press: Washington DC.
- National Center for Education Statistics. *User's Manual for the ECLS-K Longitudinal Kindergarten-First Grade Public-Use Data Files and Electronic Codebook*. Washington, DC: U.S. Department of Education.
- National Center for Education Statistics. 2003. *Early Childhood Longitudinal Survey, Kindergarten Cohort* [. Washington, DC.
- Newmann, F. M. 1991. "Student Engagement in Academic Work: Expanding the Perspective on Secondary School Effectiveness." Pp. 58-75 in *Rethinking effective schools : research and practice*. Bliss, J. R., W. A. Firestone, C. E. Richards, Eds. Englewood Cliffs, NJ: Prentice Hall.
- Raudenbush, S. W. and A. S. Bryk. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2 ed. Thousand Oaks, CA: Sage.
- Raudenbush, Stephen W. and Anthony S. Bryk. 2002b. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2 ed. Thousand Oaks, CA: Sage.
- Reardon, Sean. 2003. "Sources of Educational Inequality" Paper presented at the American Sociological Association Meetings in Atlanta, Georgia.
- Robinson, G. K. 1991. "That BLUP Is a Good Thing: The Estimation of Random Effects." *Statistical Science* 6(1):15-32.
- Rock, Donald A. and Judith M. Pollack. *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.
- Rowan, B. 1984. "Shamanistic Rituals in Effective Schools." *Issues in Education* 2:517:37. NCES 200205. Washington, DC: National Center for Education Statistics.
- Rubenstein, Ross, Leanna Stiefel, Amy Ellen Schwartz, and Hella Bel Hadj Amor. "Distinguishing Good Schools From Bad In Principle and Practice: A Comparison of

- Four Methods.” Pp. 55-70 in Fowler, W.J., Jr., ed. (2004) *Developments in School Finance: Fiscal Proceedings from the Annual State Data Conference of July 2003*, (NCES 2004-325), U.S. Department of Education, National Center for Education Statistics, Washington, DC: Government Printing office.
- Rubin, Donald B. 1987. *Multiple Imputation for Survey Nonresponse*. New York: Wiley.
- Sanders, W. 1998. “Value-Added Assessment.” *The School Administrator* 55(11): 24-32.
- Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman and Hall.
- Schafer, J. L. and R. M. Yucel. 2002. "Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values." *Journal of Computational & Graphical Statistics* 11(2):437-57.
- Singer, Judith D. and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, UK: Oxford University Press.
- Teachman, Jay. 1987. “Family Background, Educational Resources, and Educational Attainment,” *American Sociological Review* 52:548-57.
- Therstrom, Abigail and Stephan Therstrom. 2003. *No Excuses: Closing the Racial Gap in Learning*. New York, New York: Simon A& Schuster.
- von Hippel, P. T. 2004a. "Multiple Imputation With Deletion: Increasing Efficiency by Deleting Cases With Imputed Y [Working Paper]." .
- von Hippel, Paul T. 2004b. "Good News on the Accountability of Small Schools: A Comment on Kane and Staiger." *Journal of Economic Perspectives*, forthcoming.
- Walberg, Herbert J. 1984. “Families as Partners in Educational Productivity.” *Phi Delta Kappan* 65:397-400.
- West, J., K. Denton, and E. Germino-Hausken. 2000. *America’s Kindergartners: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99*, NCES 2000-070. Washington DC: U.S. Department of Education, National Center for Education Statistics.
- West, Martin R. and Paul E. Peterson. 2003. “The Politics and Practice of Accountability.” Pp. 1-20 in *No Child Left Behind: The Politics and Practice of School Accountability* (Peterson and West Eds.) Washington, DC: The Brookings Institution.
- Wilson, William Julius. 1996. *When Work Disappears: The World of the New Urban Poor*. New York: Knopf.

Table 1. Measurement error variance on reading and math tests.

Occasion	Students tested	Total variance	Reliability	Measurement error variance
<b>Reading</b>				
Fall 1998	14,441	73.62	0.93	5.15
Spring 1999	15,886	117.72	0.95	5.89
Fall 1999	4,572	160.53	0.96	6.42
Spring 2000	14,455	200.79	0.97	6.02
Average				5.76
<b>Math</b>				
Fall 1998				
Spring 1999				
Fall 1999				
Spring 2000				
Average				

Note. Reliabilities were calculated by Rock and Pollack (2002) using Item Response Theory. If the reliability is  $r$  and the total variance of a test is  $Var(Y_{sct})$ , then the measurement error variance is  $(1-r) Var(Y_{sct})$ . To average the measurement error variance across occasions, we weighted each variance by the number of students tested on each occasion.

Table 2. Schools in Bottom Quintile (“Failing”) Versus the Top Four Quintiles by Achievement, Learning, and Impact Measures. Source: Early Childhood Longitudinal Study—Kindergarten Cohort 1998

Reading		Achievement		
		Bottom quintile	Top four quintiles	
Learning	Bottom quintile	32	29	61
	Top four quintiles	29	219	248
		61	248	309

Impact		Achievement		
		Bottom quintile	Top four quintiles	
Learning	Bottom quintile	18	43	61
	Top four quintiles	43	205	248
		61	248	309

		Achievement	
		Bottom quintile	Top four quintiles
Learning	Bottom quintile	<input type="text"/>	
	Top four quintiles		

		Achievement	
		Bottom quintile	Top four quintiles
Impact	Bottom quintile	<input type="text"/>	
	Top four quintiles		

**Table 3. Average characteristics of successful and failing schools, by three different criteria.**  
**Source: Early Childhood Longitudinal Study—Kindergarten Cohort, 1998**

		Achievement		Learning		Impact	
		Bottom quintile	Top four quintiles	Bottom quintile	Top four quintiles	Bottom quintile	Top four quintiles
<b>Reading</b>							
Students	Minority	71%	32%	56%	36%	50%	37%
	Free lunch	54%	23%	36%	27%	25%	30%
	Reduced lunch	9%	7%	7%	7%	7%	7%
Sector	Public	97%	72%	75%	77%	69%	79%
	Catholic	0%	4%	13%	1%	13%	1%
	Other religious	3%	12%	7%	11%	11%	10%
	Other private	0%	12%	5%	10%	7%	10%
Location	Suburban	21%	42%	38%	38%	38%	38%
	Urban	56%	34%	41%	38%	44%	37%
	Rural	23%	24%	21%	24%	18%	25%
<b>Math</b>							
Students	Minority						
	Free lunch						
	Reduced lunch						
Sector	Public						
	Catholic						
	Other religious						
	Other private						
Location	Suburban						
	Urban						
	Rural						