

Overcoming the Volatility in School-Level Gain Scores: A New Approach to Identifying Value Added With Cross-Sectional Data

by Sean Kelly and Laura Monczunski

Traditionally, state accountability systems have measured school-level achievement gains using cross-sectional data, for example, by comparing scores of one year's eighth graders to scores of the next year's eighth graders. This approach produces extremely volatile estimates of value added from year to year. This volatility suggests that the traditional use of cross-sectional data cannot reliably estimate the production of achievement by schools, and therefore schools may be unfairly sanctioned under such a system. In this analysis, the authors consider an alternative use of cross-sectional data, identifying differences in relative subject matter performance within schools. They illustrate this approach using data on public middle schools in Wisconsin during the years 1998 to 2001. Compared to school-level gain scores, relative subject matter performance is much more stable from year to year. The authors conclude that in lieu of more reliable measures of value added, state educational agencies should consider alternative uses of standardized test data.

Keywords: accountability; educational assessment; middle schools; school effects; value added

The passage of the No Child Left Behind Act (NCLB) of 2001 solidified the growing trend toward test-based accountability. Accountability regimes are designed to increase achievement growth and promote equality of educational opportunity. With these goals in mind, schools are now identified as "in need of improvement" or failing to make "adequate yearly progress" (AYP) by a formula that individual states create, emphasizing progress toward all students being proficient on standardized tests. Actions taken against schools for failure to meet AYP in consecutive years, required by NCLB for Title I schools, escalate, culminating in reconstitution of the schools.

Because such accountability systems rely primarily on labeling and sanctioning to motivate performance, success depends on the testing programs identifying true value added by schools. Students begin each school year with a base of knowledge and academic skills. *Value added* is the additional knowledge and skills acquired during the process of schooling (Tekwe et al., 2004). The concept of value added helps distinguish between the

contribution of schooling to academic achievement and simply the observed level of achievement at a school that is a function of both school and nonschool sources. Value added is closely related to the concept of equality of educational opportunity, which is present when schooling exerts a strong influence on students that is independent of their backgrounds and general social context (Coleman, 1990). When value added (a) varies systematically across teachers or schools, (b) differs for social groups within schools, or (c) is weak compared to learning that occurs outside of schools, equality of educational opportunity is low.

The majority of state accountability programs are based on cross-sectional comparisons of different cohorts of students, for example, a comparison between the test scores of one year's fourth-grade class and the scores of the previous year's fourth graders. If a higher proportion of students meet proficiency standards in subsequent years, a school is deemed successful. Although this seems straightforward at first, such systems mostly hold schools accountable for factors beyond their control, namely, random variations in test performance and changes in the knowledge and skills that students bring to school to begin with. The evidence refuting the utility of simple cross-sectional analyses of achievement in estimating value added at the school level is quite convincing. Traditional measures of school progress are highly volatile: Schools appear to make achievement gains in one year only to lose ground the following year (Linn & Haug, 2002).

Based on the present analysis, and on four decades of school effects research, we simply do not have much confidence that state educational agencies can identify value added at the school level using cross-sectional data. Instead, we argue that accountability systems should shift the focus to within-school, rather than between-school, evaluation. We are not advocating that states abandon a commitment to promoting equality of educational opportunity between schools but simply that erroneously labeling schools as successes and failures is counterproductive to that end and should not be the goal of accountability systems.

We illustrate how data currently used for school-to-school comparisons can be more reliably applied to performance comparisons within schools. Using cross-sectional data on recent cohorts of students in Wisconsin, we examine schools' relative performances in each of four subject matter domains: math, science, reading and language arts, and social studies. Performance in each subject is compared to performance in the other subjects, such that a school is strong or weak in a given subject relative to its own

performance in other subjects. The deviation of a school's performance from that of the average school's performance (sample mean) provides a common metric for comparisons across subject matters. No evaluation is made of a school's overall level of performance. We find that relative subject matter performance is much more stable (less volatile) than aggregate gain scores. Such an analysis of relative subject matter performance could be used to identify instances of best practices within schools at the local level.

Research on School Effects: The Difficulty of Measuring Value Added

Since the landmark Equality of Educational Opportunity Study in 1965, sociologists have known that the majority of the variance in school achievement lies within, not between, schools (Coleman, 1990). Moreover, much of the observed difference in achievement between schools can actually be explained by characteristics of students rather than schools per se. Thus, it is easy to mistake factors that are really associated with students, such as differences in family background, with those of schools. In hundreds of analyses of school effects, Coleman's basic conclusions have been confirmed (Scheerens & Bosker, 1997).

To say that test scores are associated with family background should not be taken somehow as evidence of innate differences. In fact, students' capabilities to learn, as evinced by learning rates during the school year, appear to be largely independent of family background (Downey, Von Hippel, & Broh, 2004; Entwisle, Alexander, & Olson, 1997). The important insight is that because students from disadvantaged backgrounds often begin first grade with lower levels of achievement, schools serving impoverished neighborhoods can easily be wrongly identified as less effective if one focuses only on average test scores.¹

The relative weakness of school effects compared to the variation in achievement at the student level, combined with changing student demographics, contributes to the instability in average test scores at the school level from year to year.² Researchers have been aware of this phenomenon since the early 1970s (Jencks, 1972, p. 91). The passage of NCLB, and the very real consequences it attached to these yearly changes, has revived interest in investigating the stability of school-level gain scores. Thomas Kane and Douglas Staiger (2002), using data from elementary school students in North Carolina, find that 50% to 80% of observed achievement gains each year are temporary because of sampling error or other nonpersistent causes (p. 248). Robert Linn and Carolyn Haug (2002) concur that changes in school performance from year to year can be "wildly unstable." Focusing on the results of fourth-grade reading tests in Colorado, they found that the correlation between gain scores, even when averaged over adjacent years (1997–1999 vs. 1998–2000), was around $-.03$. In other words, the magnitude of the change from 1997 to 1999 indicates virtually nothing about the change from 1998 to 2000.

Growth Models of Value Added Using Longitudinal Data

Recent changes in the compilation of standardized test data in several states have allowed researchers to prepare estimates of value added based on models of achievement growth among the same group of students (Ballou, Sanders, & Wright, 2004; Raudenbush, Bryk, & Ponsciak, 2003; Tekwe et al., 2004). This approach represents a dramatic improvement over cross-sectional

approaches, but some researchers are still skeptical of the overall utility of such value-added estimates (Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004).

Although growth models overcome much of the problem of students being nonrandomly matched with schools, several challenges remain. Because growth models mostly rely on data from multiple school years, they are susceptible to unreliability due to student mobility and missing data (Rubin et al., 2004).³ Using time points from separate years also confounds achievement growth during the school year with achievement growth during the summer, a known source of bias (Downey et al., 2004; Entwisle et al., 1997). The overall precision of achievement growth models is often low. At the teacher level, only a small fraction of individuals can be reliably identified as above or below average (Ballou et al., 2004). Even at the school level, Raudenbush (2004) argues that value-added estimates are probably best averaged over multiple years.

Aggregating data to higher levels, from teachers or classrooms to the school level and from individual subjects to average performance, increases the reliability of the value-added estimates somewhat by averaging out sources of error (Linn, Baker, & Betebenner, 2002; Raudenbush, 2004). Unfortunately though, at each level of aggregation the connection to classroom instruction becomes one step further removed. Moreover, as statistical models become increasingly complex, educational practitioners might find the value-added estimates of their classrooms' or schools' performances to be too abstract to appreciate. The simplest yet adequate approach to measuring value added is preferred (Tekwe et al., 2004).

We believe many of these difficulties can be overcome. Researchers should continue to develop reliable and parsimonious growth models of value added, and state educational agencies should support these efforts by collecting and organizing the necessary longitudinal data. In the meantime, we advocate a dramatic shift in the use of cross-sectional standardized test data, from comparisons across schools, which are known to be unreliable, to potentially useful within-school analyses. The relative subject matter performance comparison we investigate in this analysis has several desirable properties. First, and most important, for each school relative subject matter comparisons are computed using the same set of students. Traditional comparisons are across years on different sets of students, which could account substantially for changes in performance from year to year. Second, because a relative subject matter analysis is specific to individual school years, it is not as greatly affected by student mobility as multiyear comparisons are. Third, the analysis is aggregated across teachers and classrooms, improving the reliability of the subject matter performance estimates, but it is not so aggregated that informing instructional improvement becomes difficult. Finally, it does not rely on an overly complex statistical procedure. Table 1 provides a comparison of growth models (the traditional approach) and relative subject matter indicators. The comparison is qualitative in nature and reflects our best estimation, not established empirical evidence.

Our analysis consists of three parts. We begin by presenting some background information regarding education in Wisconsin and particularly on 8th-grade achievement during this period, which is the grade level of data investigated. Next, we replicate Linn and Haug's (2002) findings on the instability of school-level

Table 1
Proposed Properties of Three Approaches to Identifying Value Added

Approach to Identifying Value Added	Properties Affecting Reliability			Properties Affecting Use in Test-Based Reform	
	Problems of Student Mobility and Missing Data	Nonrandom Matching of Students to Schools (Selection Bias)	Level of Aggregation: Subject Matter, Grades, or Years ^a	Statistical Complexity	Between-School Comparisons
Traditional approach: cross-sectional comparisons of subsequent cohorts	Very high ^b	Very high	Variable	Low	Yes
Growth models using longitudinal data	High	Very low	Variable	Very high	Yes
Relative subject matter performance	Low	Low	Low	Low	No

^aLevel of aggregation affects both statistical reliability and use in test-based reform.

^b*Not applicable* would also be an appropriate term because the traditional approach relies solely on comparisons of different students.

gain scores. Finally, we present the relative subject matter performance analysis. Throughout, we reference Harrison Middle School, which tends to hover around many of the state averages, to illustrate our analysis.⁴

Data and Method

We analyze Wisconsin Knowledge and Concepts Examinations (WKCE) data for the years 1998 to 2001 for eighth-grade students in Wisconsin. These data are available to the public in aggregated form on the Wisconsin Department of Public Instruction website. The WKCE, developed by Wisconsin educators together with CTB/McGraw-Hill, includes questions derived from the Wisconsin Model Academic Standards and questions used in standardized tests nationwide. In 2005, Wisconsin switched to a new version of this test (WKCE–Criterion Referenced Test), which is exclusive to the state. However, for the years we analyze, the WKCE was essentially a customized version of TerraNova, a nationally normed test. Each fall, students in Grades 3 through 8 and 10 are tested on reading and mathematics (in accordance with NCLB). In addition, students in Grades 4, 8, and 10 are tested on science, social studies, and language arts. The Wisconsin Model Academic Standards includes 14 other subjects, but these are not directly tested. Although schools have been required to consider WKCE scores when making grade promotion decisions for fourth and eighth graders since 2002, high stakes for individual students are not attached to the test in most cases because other criteria, such as grades and teacher recommendations, factor heavily in these decisions as well.

We focus on middle school students because they have different teachers for different subjects, which we believe is a source of differential value added within schools in our analysis. The data set included 620 schools, but because of missing data most of the analysis uses 479 schools. The number of students in the data set varied from a low of 66,238 in 1998 to a high of 68,123 in 1999. The number of students in the 479 schools ranged from a low of 57,385 in 2001 to a high of 58,985 in 1999. The mean enrollment for Wisconsin's eighth-grade classes was between 108 and 114 during 1998–2001, with enrollments ranging from 1 student to 428 students. Harrison Middle School, the example school, had an eighth-grade enrollment ranging from 68 to 75, which means

that a student there would be placed in one of four or five classrooms of eighth graders.

We utilize an approximate student fixed-effects design. That is, by examining test scores of different subjects within schools, we are comparing scores largely of the same set of students. This allows us to be better able to determine the value added by a school, because influences on achievement associated with students will not vary much for the scores that we compare.⁵ For each of the four subject matters, we computed the ratio of the proportion of students scoring at the proficient or advanced category in each school to that of the sample mean of the other 478 schools in the sample. We then computed a similar ratio of the school's performance to that of the sample mean in the other three subjects combined. The relative performance in each of the four subject matter areas is calculated as the difference between the two calculations, one specific to a given subject, the other pertaining to the remaining subjects. This produces a value for the relative subject matter performance for each subject within each school. We then estimate the stability of performance in each subject relative to the other subjects over time. We utilize the percentage proficient/advanced, which is publicly available data. However, we recommend that such an analysis be conducted using raw scores (scale scores) whenever these data are available.⁶

Results

Overall, students in Wisconsin's schools compare favorably with students in other states on achievement tests. In 1998, according to the National Assessment of Educational Progress, reading scores among Wisconsin students in both fourth and eighth grades surpassed those of the national average (National Center for Education Statistics, 1999). For example, among fourth graders, 34% of Wisconsin students were reading at or above the proficient level, whereas the national average was at 29%. At the lower end of the ability distribution, only 28% of Wisconsin fourth graders were below the basic-level cutoff, compared with 39% in the nation as a whole.

These results are not surprising when one considers that Wisconsin is a relatively affluent state. Because school achievement is correlated with a student's social class, this factor is likely to be an important component of Wisconsin's overall

performance (Downey et al., 2004; Jencks, 1972). Of the 877,753 public students enrolled in Wisconsin during the 1999–2000 school year, only 25% were eligible for free or reduced lunch. Only Massachusetts, Vermont, and New Hampshire have fewer students involved in these programs. Likewise, in 1998–1999, Wisconsin spent \$7,527 per pupil on education, placing it far ahead of most states and every mid-western state with the exception of Michigan. Although the link between school expenditures and achievement is not particularly strong, it certainly affects the resources available to students (Greenwald, Hedges, & Laine, 1996). For example in 1999–2000, the ratio of 14.4 students per teacher in Wisconsin was lower than in the nation as a whole, which averages 16.1 (National Center for Education Statistics, 2001).

Yet within Wisconsin's borders, economic and social resources are not evenly distributed, and this affects the distribution of achievement levels across Wisconsin schools. Take, for example, Milwaukee County, the most populous county in Wisconsin. Historically the base of Wisconsin's strong manufacturing economy, until recently Milwaukee has always had more than its fair share of high-paying blue-collar jobs. In recent decades though, Milwaukee has developed an alarming urban poverty problem. Like in other urban areas, during the 1980s the geographic concentration of poverty accelerated in Milwaukee. The changing economy had a disproportionately negative impact on Milwaukee's Black residents. Paul Jargowsky (1997) describes how by 1990 almost half of all of Milwaukee's Black residents were living in ghetto neighborhoods, or in predominantly Black neighborhoods where more than 40% of the population was living in poverty, as opposed to only 16% in 1970. The highly segregated neighborhoods of Milwaukee map onto an equally segregated school system. In 1999, the average Black student in Milwaukee attended a school that was 78.2% Black (Lewis Mumford Center, 2002).

In Wisconsin, as elsewhere in the nation, the system of school accountability forced into place by NCLB is a bit of a farce. If students were randomly assigned to schools, then we might be able to take average test scores as evidence of a school's effectiveness. But students are not allocated randomly across schools, and in any given year average school performance primarily reflects the achievement levels of students entering schools, not large differences in value added across schools. Again, this conclusion is not about students' innate capabilities or ultimate possibilities for educational success. Rather, it acknowledges that eighth-grade teachers whose students begin the school year reading at the fifth-grade level have a more difficult task ahead of them than teachers whose students begin the year at the seventh-grade level.

Wisconsin Eighth-Grade Achievement: 1998–2001

After taking the WKCE, each student is assigned to one of four categories—minimum, basic, proficient, or advanced—based on his or her score. Figures 1 and 2 illustrate the mean percentages of students in each category in each year from 1998 to 2001 in reading and mathematics (other subjects are reported numerically in Table 2). The proficient and advanced categories are combined because NCLB holds schools accountable only for meeting the level of proficient. Proportions in the proficient/advanced

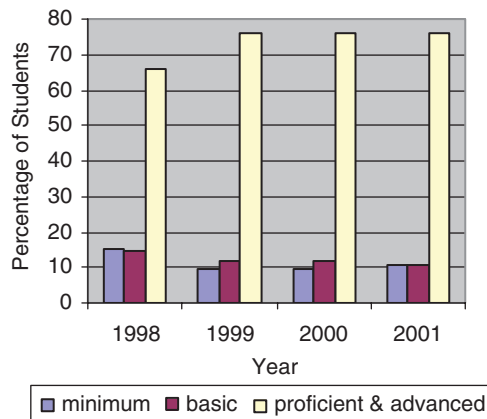


FIGURE 1. Reading test mean percentages.

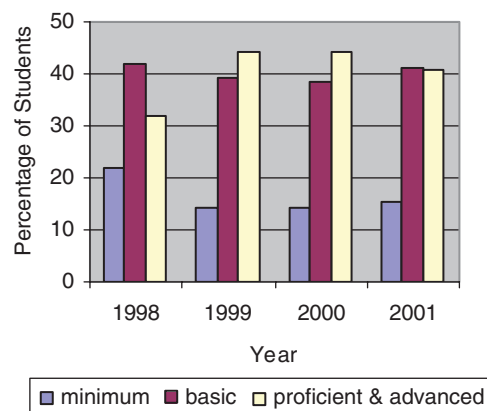


FIGURE 2. Math test mean percentages.

category generally increased in the beginning of the 4-year period, followed by a slight decrease. This decrease was small enough that they still finished at slightly higher levels than those at which they began. The only exception to this trend is the language test, which shows a large gain in scores from 1998 to 1999 (see Table 2). This dramatic change was caused by a change in the test itself—language arts followed the general trend in the remaining 3 years.

Table 2 displays the precise numerical values for the mean percentage of students at proficient/advanced levels, along with the correlation between the schools' initial levels of achievement and their 4-year gains. If the principal at Harrison looked at these data, she or he would see a level of performance close to the state mean for reading. The principal would aim at improving, but in the context of generally stable or even improving performance. The mean percentages proficient/advanced in the sample as a whole for reading in chronological order were 65.72%, 75.97%, 75.97%, and 75.86%. Harrison Middle School's percentages were about 2% lower in the 1st year, about 3% higher in Years 2 and 3, and about 4% lower in Year 4. Harrison's other test scores follow the general up and down pattern seen in Figures 1 and 2 and Table 2.

Correlations (shown in the bottom row of Table 2) between schools' initial levels of achievement and schools' 4-year gains

Table 2
Initial Level of Proficient/Advanced and Correlation Between Level and 4-Year Gain:
Wisconsin Eighth Graders, 1998–2001

	Reading	Reading—Harrison	Language Arts ^a	Science	Social Studies	Math
1998	65.72	63	18.74	58.97	72.14	31.80
1999	75.97	79	73.93	69.83	85.21	44.26
2000	75.97	79	73.43	70.86	83.68	44.22
2001	75.86	72	66.84	65.25	82.71	40.89
$r(1, 1-4)$	-.334**	n/a		-.256**	-.229**	-.306**
$r(2, 2-4)$			-.229**			

^a $r(2, 2-4)$ is used for language arts due to testing changes from 1998 to 1999.

** $p < .001$.

show that differences in performance between Wisconsin schools are evening out. Regression to the mean is occurring: Higher performing schools and lower performing schools are converging to the same level of achievement as initially lower performing schools make more substantial gains. For instance, for the reading test $r(\text{proficient/advanced } 1998, \text{ gain } 1998 \text{ to } 2001) = -.34$. However, we cannot say what is driving this process—changes in value added by schools or perhaps changing demographics, or maybe it is simply a statistical artifact of correlated errors. The magnitude of the reading correlation is largest, most strongly indicating regression to the mean, but science, social studies, and math follow the same pattern.

Linn and Haug Replication

Currently, the majority of accountability systems are based on comparisons of different cohorts of students across time. Table 3 demonstrates how one could initially believe gain scores to be a relevant method of measuring progress and why this would be a mistaken belief. The correlations between the gain in the 1st year and the 4-year gain are fairly high, seeming to indicate that they would be a useful tool in assessing schools. However, when we remove the shared year from the correlation, and instead look at the gain in the 1st year and the gain in the past 3 years, the correlations become negative. If these data were well suited to identifying value added there would not be such a radical difference between $r(1-2, 1-4)$ and $r(1-2, 2-4)$.⁷

The final column in the table uses 2-year average gain scores (Linn & Haug, 2002). For instance, the average of the gain in Year 1 to 2 and the gain in Year 2 to 3 is written as 1–3 in Table 3. In reading, the mean average gain from Year 1 to 3 was about 5.13 and from Year 2 to 4 was about -0.052 . Harrison Middle School followed the same pattern but with somewhat larger changes—a gain of 8 and a loss of 3.5. When we use 2-year average gain scores, the correlations become small enough that they are insignificant; that is, the average gain from Year 1 to 3 tells almost nothing about the average gain from Year 2 to 4. Figure 3 is a scatter plot of the relationship between these two gain scores for reading. This visually demonstrates the weak correlation found in Table 3—the points are grouped in a cloud with no clear up or down trend.

One could question whether the instability seen in these correlations is caused by several outlier schools with unusually high gains or losses. To test this, we constructed the same correlations, $r(1-3, 2-4)$, but this time omitted schools that experienced gains

Table 3
Correlations in School-Level Gain Scores:
Wisconsin Eighth Graders, 1998–2001

Subject	$r(1-2, 1-4)$	$r(1-2, 2-4)$	$r(1-3, 2-4)$
Reading	.568	-.467	.049
Language arts	.719	-.339	.022
Science	.523	-.450	.044
Social studies	.628	-.415	.003
Math	.545	-.381	.116*

Note. All correlations use the same 479 schools.

* $p < .05$.

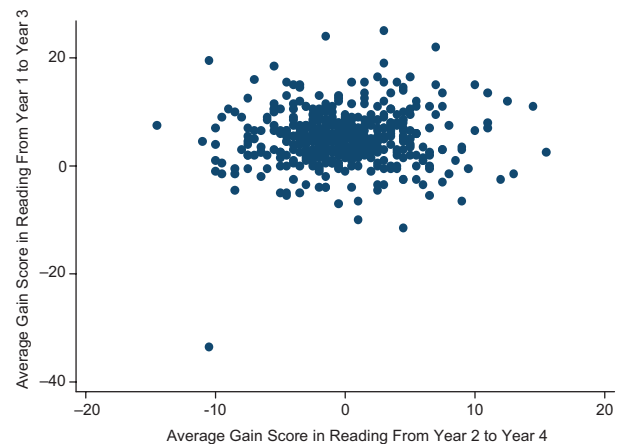


FIGURE 3. Scatterplot of school-level reading gain scores in adjacent 2-year periods.

or losses larger than 15% proficient/advanced. Performing this analysis has practically no effect on the correlations, which indicates that a few outliers are not the cause of the instability. Other supplementary analyses confirmed the results in Table 3.⁸ Although the correlations are positive (they were actually negative in Linn & Haug’s 2002 analysis), these data confirm Linn and Haug’s findings of the instability of gain scores derived from cross-sectional data. Only in mathematics is there any stability whatsoever ($p < .05$).

The Stability of Relative Subject Matter Performance Within Schools

To analyze relative subject matter performance within schools we first calculated ratios describing the performance of each school in a given subject to that of the sample as a whole. A school with a ratio of 1.01 in math is about average scoring in math compared to other Wisconsin schools, whereas values less than 1 indicate weaker performance, and values greater than 1 indicate stronger performance. Next, to identify relative subject matter performance within schools, for each year we took the ratio of performance in one subject and subtracted the average performance of all other subjects from it. This means that a school with a relative subject matter performance of .1 in math scored better in math than it did in the other subjects on average.

Table 4 shows the three-step calculation of relative subject matter performance for Harrison Middle School in 1998. The first column (A) is how well Harrison performed in each subject compared to other Wisconsin schools. The second column (B) is the average of Harrison’s performance in each subject except the one in that particular row. For instance in the reading and language arts row, 1.04 indicates Harrison’s performance in science, social studies, and math compared to other schools. Finally, the last column shows the relative performance of each subject at Harrison, which is simply column A-B. The differences across subjects at Harrison reported in Table 4 are quite large. In 1998, science was a strong subject at Harrison. The ratio of 1.12 might be understood by comparing how Harrison ranked in that subject compared to other schools—the average rank position was a full 123 places ahead of the average rank in other subjects (Harrison ranked 150th in science but only 273rd in the other subjects on average). If science was a strong subject at Harrison, reading and language arts stands out as the lowest performing subject for Harrison, falling substantially behind the school average.

Figure 4 is a histogram of relative performance in reading for 1998. It is typical of histograms for other subjects and other years. Each column encompasses a .02 difference, or about 15 ranking spots, and the numbers at the top of each column represent how many schools’ relative performance in reading falls in that column’s .02 unit range. The distribution is balanced around zero (which is a function of the definition/calculation of relative performance). All schools represented in the bars to the right of zero

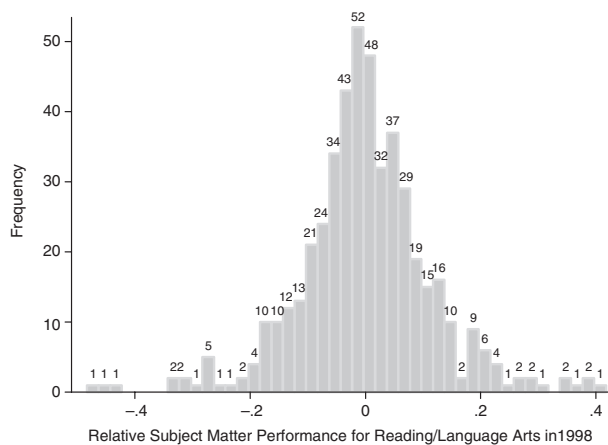


FIGURE 4. Histogram of relative subject matter performance for reading and language arts in 1998.

were performing better in reading than their average. The best school on this histogram had a relative reading ratio .4 units higher than its performance in other subjects, which translates into more than 283 places better in rank position in reading than in other subjects. Schools to the left of zero performed worse in that subject. Harrison falls fairly far to the left on the reading histogram, with a deviation of $-.13$. A student at Harrison that year most likely learned substantially less from the reading instruction than from instruction in other subjects.

Correlations between relative subject matter performance (see Table 5) in adjacent years and 2-year averages show that relative performance tends to be somewhat stable from year to year and much more stable than simple gain scores based on the performance of different groups of students. The adjacent year correlations are between .23 and .53. The correlations using 2-year averages are stronger still (from .398 in science to .626 in social studies). Compared to the stability of school-level gain scores, relative subject matter performance is at least 5 to 10 times as stable. Tests of statistical significance of $r(1-2, 3-4)$ indicate that relative subject matter performance has some stability in all subjects (the correlation is positive and greater than zero). The coefficients in Table 5 indicate that cross-sectional data can

Table 4
Relative Subject Matter Performance at Harrison Middle School in 1998

	(A) Ratio of Harrison's Performance in Individual Subjects to Sample Mean	(B) Ratio of Harrison's Performance in Other Subjects to Sample Mean ^a	(C) Relative Subject Matter Performance (A-B)
Reading and language arts	0.91	1.04	-0.13
Science	1.12	0.97	0.15
Social studies	1.00	1.02	-0.02
Math	0.97	1.02	-0.04

^aThe average ratio of all four subjects at Harrison is 1.01.

Table 5
Stability of Relative Subject Matter Performance in
Each Subject at the School Level: Wisconsin Eighth
Graders, 1998–2001

	<i>r</i> (1, 2)	<i>r</i> (2, 3)	<i>r</i> (3, 4)	<i>r</i> (1–2, 3–4)
Reading and language arts	.232	.302	.333	.408
Science	.325	.277	.326	.398
Social studies	.532	.459	.490	.626
Math	.504	.497	.537	.573

Note. For all correlations, $p < .001$.

provide meaningful information, at least to some schools with large differences in performance across subject matter domains.

Discussion

How might the analysis of relative subject matter performance within schools inform educational policy and practice? Current use of standardized test data driven by NCLB emphasizes external accountability—schools are labeled as high and low performing and much of the “solution” to the problem of chronically low performing schools is put in the hands of parents and state educational agencies. But standardized test data can also be a valuable tool internally, providing information about student performance that schools can use to improve their curriculum and instruction. Like Linn (2003), we believe that educational policy should shift focus from using tests to label schools as failures to using tests to inform improvements in schools. California’s Best Practices initiative provides an example of such an approach by using standardized test data to identify successful schools. A geographically balanced mix of schools that scored well on the California Standards tests, after adjusting for demographic composition of the student body, are selected for case study (Springboard Schools, 2007). Best practices from selected schools at the elementary, middle, and high school are then disseminated statewide in research reports and periodicals (e.g., the California Department of Education periodical *High School!*). Measures of relative subject matter performance could be a valuable tool in identifying successful classroom instruction.

Consider Harrison Middle School. Averaging over all subjects, it is about at the mean achievement level for the state as a whole. In 1998, the level of achievement in science was quite high, whereas it was weak in reading and language arts. Three years later, in 2001, the pattern was similar, with science becoming the second-best subject and reading and language arts remaining the worst. The staff at Harrison might begin to consider the nature of instruction in their science classrooms more carefully. What makes these classrooms so successful with the same students that other teachers are less successful with? The science teachers at Harrison may be a professional development resource for the whole school. Another way to address differential performance might be to integrate instruction in skills from weaker subjects into subjects where students are already highly engaged and learning rapidly. For example, Harrison’s teachers might address the low performance in English by placing more emphasis on literacy skills in the context of students’ science classes.

Schools are struggling to cope with the consequences of NCLB. Under the current accountability regime, which in most states is based on simple mean proficiency scores in cross-sectional data, a steadily improving school with a high poverty rate may easily be labeled as failing because of selection bias (Kim & Sunderman, 2005). In addition, schools with more subgroups are less likely to meet AYP, and in this way NCLB unintentionally favors schools with fewer minority populations. Moreover, teachers and administration may be responding to the mounting pressure in ways that are counterproductive to widespread educational success (Booher-Jennings, 2005). Teachers in the school Booher-Jennings observed engaged in forms of “educational triage,” focusing instructional efforts on those “bubble students” who would make the most difference in improving the proportion of students meeting proficiency requirements. One teacher articulated that they must focus their efforts on certain students close to the proficiency standard, as they cannot afford to spend time with “remedial kids,” who are a “lost cause” (Booher-Jennings, 2005, p. 241). Because of the performance benchmark Wisconsin uses for AYP calculations, only a small fraction of Wisconsin schools were at risk of not making AYP in 2004–2005 (2%) and 2005–2006 (4%), and strong accommodations of the type Booher-Jennings found were probably rare. But in other states, such as Hawaii, Florida, Rhode Island, Nevada, New Mexico, South Carolina, and the District of Columbia, 50% or more of all schools failed to meet AYP in 2004–2005 (Olson, 2006).

We must change our approach to school accountability in the United States. The current emphasis on labeling schools as high performing and low performing might make sense under a system where these labels were reliable. Under our current system, we cannot even address this question because most states do not have a system in place to reliably measure school performance. The analysis in this study suggests one way to refocus school accountability: move the focus of school accountability from between schools to within schools.

Perhaps the risks inherent in focusing on relative subject matter performance, such as negative effects on teacher collegiality or teachers’ sense of autonomy, would outweigh any benefits of such an approach. The potential negative effects of such an approach need to be carefully studied. It is also important to reiterate that relative subject matter analyses say nothing about schools as a whole. First, we hope that an emphasis on within-school achievement comparisons would not distract from efforts to address between-school inequalities in achievement and to ensure that all schools have the resources they need to be successful, including a qualified faculty (Lankford, Loeb, & Wyckoff, 2002). Second, we acknowledge that under such an approach, a school that is failing its students in all subjects would look similar to a school that is extremely successful in all subjects. This is an important criticism. Hopefully, new developments in value-added modeling (e.g. Meyer, 2007) will yield measures of performance that can distinguish between schools with effective and ineffective instruction, and states will do what they can, given limited resources for additional data collection, to embrace those developments.

In the meantime, existing cross-sectional data on school average levels of performance can still be put to use identifying schools that need additional resources and highlighting examples of best practice. An important strength of the relative subject matter performance analysis described here is that it does provide

a synopsis of achievement within schools that has some relevance to actual value added. Perhaps if such an approach were emphasized, schools and districts would be able to concentrate on bolstering instruction in subject matters in need of improvement and acknowledging legitimate excellence, instead of hunting for ways to meet the bottom line of an accountability system that is based on essentially meaningless information.

NOTES

This research is made possible in part by support from the Institute for Scholarship in the Liberal Arts, College of Arts and Letters, University of Notre Dame.

¹Entwisle, Alexander, and Olson (1997) did in fact find that students in schools of low socioeconomic status were erroneously treated as if they were less capable, with much higher rates of retention and special education placement, despite average or even above average achievement growth during the school year (see chap. 4).

²Year-to-year changes in student demographics could certainly account for the small changes in average test scores from year to year. Tekwe et al. (2004) report student demographics in the third and fifth grades for 19 schools in their value-added analysis. The average change in percentage free-reduced lunch was 7.3 percentiles, with a maximum difference of 19.1 percentiles. Models including student demographics (#3 HLMM) produced significantly different value-added estimates from models without student demographics (#1, 3, 4). Researchers lack the data to test this hypothesis explicitly because databases that contain detailed data on family background and other student characteristics are typically longitudinal in nature (not consecutive cross sections) and have too few students sampled within schools to produce reliable estimates of school average achievement levels (e.g., the National Educational Longitudinal Survey)

³Alternately, data from several time points during the school year could be used. This would accomplish the goal of separating growth from initial achievement while reducing the time span during which mobility can occur. Unfortunately, that would also place a greater testing burden on teachers during a single year.

⁴Although all data used in this analysis are publicly available, Harrison is a pseudonym. In addition, the achievement data for Harrison has been slightly obscured by minor deviations.

⁵Subject matter domains of the Wisconsin Knowledge and Concepts Examinations (WKCE) were administered on different days, and this resulted in a small amount of nonoverlap in the student populations taking tests across subjects. However, because schools are encouraged to conduct makeup tests within the 5-week testing window, there was on average less than a 1% difference in the students taking different components of the WKCE.

⁶We used STATA (version 8.2) software to analyze the data.

⁷The variability in gain scores across adjacent years was similar in these data; thus each 1-year gain contributed about equally to the variance in the pooled gain scores (1–4, 2–4, etc.). The correlations in Table 3 were not unduly affected by the large increase in test scores from 1998 to 1999.

⁸We performed an analysis of gain scores excluding schools with enrollments of fewer than 20 from our sample. The results were almost identical to the analysis with all schools—small schools behaved similarly to larger schools. In addition, we conducted an analysis to determine whether the data were experiencing ceiling effects by excluding schools with 90% or more of their students already at the proficient/advanced level from the analysis (we found that it was not).

REFERENCES

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29, 37–65.

- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal*, 42, 231–268.
- Coleman, J. S. (1990). *Equality and achievement in education*. Boulder, CO: Westview.
- Downey, D. B., Von Hippel, P. T., & Broh, B. (2004). Are schools the great equalizer? *American Sociological Review*, 69, 613–635.
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, schools, and inequality*. Boulder, CO: Westview.
- Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, 66, 361–396.
- Jargowsky, P. A. (1997). *Poverty and place: Ghettos, barrios, and the American city*. New York: Russell Sage.
- Jencks, C. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Harper Colophon.
- Kane, T. J., & Staiger, D. O. (2002). *Volatility in school test scores: Implications for test-based accountability systems* (Brookings papers on education policy). Washington, DC: Brookings Institution Press.
- Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher*, 34(8), 3–13.
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, 24, 37–62.
- Lewis Mumford Center. (2002, March 29). *Choosing segregation: Racial imbalance in American public schools, 1990–2000*. Retrieved April 19, 2002, <http://mumford.albany.edu/census/report.html>
- Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32(7), 3–13.
- Linn, R. L., Baker, E. L., & Betebenner, D. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31(6), 3–16.
- Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24, 29–36.
- Meyer, R. (2007). *Center for Value-Added Research at Milwaukee Public Schools*. Retrieved June 29, 2006, http://www.wcer.wisc.edu/projects/projects.php?project_num=219
- National Center for Education Statistics. (1999). *The NAEP 1998 reading report card: National and state highlights* (NCES No. 1999-479). Washington, DC: U.S. Department of Education.
- National Center for Education Statistics. (2001). *The digest of education statistics, 2001* (NCES No. 2002-130). Washington, DC: U.S. Department of Education.
- Olson, L. (2006). As AYP bar rises, more schools fail. *Education Week*, 26(4), 1, 20–21, 23.
- Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29, 121–129.
- Raudenbush, S. W., Bryk, A. S., & Ponsiaki, A. (2003, April). *School accountability*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29, 103–116.
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. New York: Pergamon.
- Springboard Schools. (2007). *California best practices study*. Retrieved March 23, 2007, from http://www.springboardschools.org/research/best_practices.html
- Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29, 11–36.

AUTHORS

SEAN KELLY is an assistant professor of sociology at the University of Notre Dame and the Center for Research on Educational Opportunity, 1015 Flanner Hall, University of Notre Dame, Notre Dame, IN 46556-5611; *kelly.206@nd.edu*. His research has focused on several educational issues facing America's schools, including problems of student engagement, the process of matching teachers to classrooms, the assignment of diverse students to course sequences in high school, and the causes of teacher attrition.

LAURA MONCZUNSKI is a graduate of the University of Notre Dame and a graduate student in the Department of Speech, Language, and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, IN 437907-2038; *lmonczun@purdue.edu*.

Manuscript received November 9, 2006

Revisions received April 17, 2007, and June 25, 2007

Accepted July 2, 2007