Journal of Research in Rural Education, 2005, *20*(13)

Random Variation in Student Performance by Class Size: Implications of NCLB in Rural Pennsylvania

Stephan J. Goetz
The Pennsylvania State University

Citation: Goetz, S. J. (2005, December 22). Random variation in student performance by class size: Implications of NCLB in rural Pennsylvania. *Journal of Research in Rural Education*, *20*(13). Retrieved [date] from http://jrre.psu.edu/articles/20-13.pdf

Schools that fail to make "adequate yearly progress" under NCLB face sanctions and may lose students to other schools. In smaller schools, random yearly variation in innate student ability and behavior can cause changes in scores that are beyond the influence of teachers. This study examines changes in reading and math scores across Pennsylvania's schools over time. There is no evidence that rural or smaller schools are systematically disadvantaged by NCLB. In smaller schools, 80% of the increase in scores is estimated to be caused by factors that generally cannot be influenced by school staff. Some schools likely should have received an award in a given year, but failed to earn one because nonpersistent factors reduced their scores to a level that disqualified them. Poverty depresses the gains achieved by schools, but small schools are able to offset the negative effect of poverty on changes in scores over time. Also, increases in poverty in a district reduce the gains in scores that can be achieved.

The No Child Left Behind Act of 2001 (NCLB) profoundly changed the Elementary and Secondary Education Act of 1965 by increasing the federal government's role in K-12 education. Beginning in 2005-2006, schools are required to test students annually in grades 3-8 in math and reading and once in grades 10-12, and they will be held accountable for their students' academic progress. Schools that continually fail to make "adequate yearly progress" on these annual assessments will face sanctions.

A potential concern with NCLB is that school size can affect average academic scores in any one year. More specifically, smaller schools have fewer students in each classroom cohort. In this case, random variation from one year to the next in innate student abilities and behaviors can lead to changes in scores at each grade level that are completely beyond the influence of teachers and administrators (Goetz & Debertin, 1991). Under these circumstances, rewards or penalties will not reflect accurately the actual performance of teachers.

The present study examines reading and math scores for grade 8 across the schools of Pennsylvania's 67 counties, for the school years 1997-1998 through 2001-2002. Pennsylvania has the fourth largest rural population, as well as one of the most populated metro areas in the U.S. (Philadelphia, with 1.5 million residents). The number of pupils in eighth grade varies across schools from fewer than 20 to 668. In addition, economic conditions vary dramatically across the different counties, providing a rich laboratory for examining determinants of student performance that are beyond the control of teachers. In rural districts, with smaller schools, it is potentially important also to consider the effects of local economic dislocation, such as the loss of manufacturing plants, or high rates of in- and out-migration on student performance. One school principal relates the story of a student who had moved into his school district from another state, then moved to an adjacent district only to move back to the original state, and then to move back into his district—all in the span of one school year. In this case, it is difficult to see how the student's family situation would not affect the student's performance on tests at the end of the school year. No allowance is made in the reporting of tests for this kind of transient behavior of students.

Loss of a major employer is likely to induce population out-migration. Workers with opportunities elsewhere are

---

This article is one of several in a special collection for which Lionel J. Beaulieu and Robert M. Gibbs served as guest editors (http://www.umaine.edu/jrre/20-12.pdf). Earlier versions of these articles were presented at the 2003 conference, "Promoting the Economic and Social vitality of Rural America: The Role of Education," which was sponsored by the Southern Rural Development Center, the Economic Research Service of the USDA, and the Rural School and Community Trust.

Correspondence concerning this article should be addressed to Stephan Goetz, The Pennsylvania State University, Agricultural Economics and Rural Sociology, 7-E Armsby, University Park, PA 16802. (sgoetz@psu.edu)

the most likely to leave a depressed area, and these may be the parents who most encourage their children to perform well in school. To the extent that economic dislocation, and other problems such as poverty are beyond the influence of school district superintendents, it may not be "fair" to hold schools accountable without controlling for the influence of these forces on student performance.

Background on School Performance Funding in Pennsylvania

The genesis of School Performance Funding (SPF) in Pennsylvania was a recommendation from the Governor's Advisory Commission on Public School Finance in 1996 (Pennsylvania Department. of Education, 2000, p. i):

> The relationship between public funding for education and the performance of students, schools and school districts must be given greater weight in the public school finance system.

SPF is intended to help "all citizens in evaluating a public school's qualities." The requirements for SPF are laid out in Pennsylvania Act 49 of 1998.

Philosophically, SPF seeks to reward the "human traits" of achievement—measured by the Pennsylvania System of School Assessment (PSSA) math and reading scores in grades 5, 8, and 11—as well as effort, which is based on average daily attendance rates. Schools need to raise their combined scaled math and reading scores by 50 points to qualify for an award, and improve school attendance rates by at least 0.75 percentage points. Further, the SPF system recognizes that schools face an implicit ceiling for scaled scores, and that it may be more difficult for the higher-performing schools to improve continually their performance. Schools that maintain high baseline scores for 3 consecutive years (minimum combined score of 2,850 in 2002) are eligible for a "Maintenance of High Standards" (MHS) award.

In addition, the state has a Governor's Achievement Schools award, as well as a Closing the Academic Achievement Gap award. The former award recognizes large improvements in test scores over consecutive years. The latter award provides for a reward multiplier of 2 if the combined score is below 2,300, and a multiplier of 1.5 if it is between 2,301 and 2,400. Thus, the lower-performing schools have an additional incentive to raise their scores, with the idea that the additional funds will over time help those schools become better performers. In this context it is noteworthy that extra awards are given to top and bottom performers, but not explicitly those with high poverty that are doing well.

The amount of funds available under SPF has increased steadily since the 1997-1998 school year ($10.4 million) to 16.8 million in 1999-2000 and $25.0 million in 2001-2002. The amount awarded is based on the size of the improvement as well as the number of students enrolled in the school. In 2002, 878 schools received an award, which is granted subject to a number of restrictions in terms of how it can be used.

In principle, schools compete only against themselves when they pursue an award because they have a target improvement level of 50 points on the scaled math and reading scores regardless of their current score level. After receiving an award, the score attained becomes the new baseline, which is frozen for 6 years. If the baseline is not exceeded after 6 years, it is reestablished based on the scores from the prior 2 years.

Measuring Student Performance: Options and Issues

States basically have available three options for measuring student performance on tests, in addition to using other measures such as attendance rates. The first is the actual level of the scale test score, which often is reported as the percentage of students falling into different achievement ranges (e.g., failing, passing, top-scoring, or novice and apprentice). A second option is to focus on changes in scores over time. This is adopted in Pennsylvania, where schools need to improve their math and reading scores by a total of 50 points. One possibility is to expect higher gains from lower-performing schools than from those that are already near or at the top. A third possibility is the use of gains in achievement either of individual students or by grade level. Tracking individual students is more expensive, while measuring the value-added over a 4-year period between fifth and eighth grade, for example, is more straightforward but also less accurate to the extent that the student population changes over 3 years.

The last two measures—changes in scores over time and value-added between different grade levels –compare schools against their own past performances. In particular, they tend to hold constant the student population and its demographic background. However, as Kane and Staiger (2002a) note, these two variables are also more difficult to measure accurately.

A critical issue in all high-stakes accountability schemes is the distinction between factors that school leaders can and cannot control in terms of their students' performance. For example, superintendents can control to some degree teacher quality and the curriculum used, but not how a particular curriculum interacts with the testing system at their particular school. In addition, superintendents have limited, if any, influence over local crime rates, school size, disruptive students, economic conditions, parental education, distances over which students are transported, or the weather on the day(s) of the test. This raises the fundamental question of whether it is fair to hold school leaders accountable for all of these factors.

Table 1
PSSA Scores, Summary Statistics: 1997-1998, Grade 8

686 Schools		498 Districts	
Reading *M* (*SD*)	Math *M* (*SD*)	Reading *M* (*SD*)	Math *M* (*SD*)
1300.6 (93.6)	1305.5 (87.9)	1322.0 (60.0)	1322.9 (65.9)

Source: Author's calculations (unweighted data).

Smaller schools also inherently exhibit greater variation in test scores over time. This variation is not just due to greater heterogeneity of small schools, since change scores are also more variable among smaller schools, as we will see shortly. In particular, the variation in scores is smallest for districts, then schools, and largest for individual students (Coleman et al., 1966).

Data Sources, Units of Analysis, and Summary Statistics

The analysis presented here is based on PSSA math and reading scores, available at http://www.paprofiles.org/pa0001/archives.htm (see also Pennsylvania Department of Education, 2002). Primarily data from grade 8 are used, although grade 5 scores are used to calculate the value-added estimates. The time span covered includes the 1997-1998, 1998-1999, 1999-2000, 2000-2001 and 2001-2002 school years. Finally, the analysis is carried out at both the school- and district-levels.

Table 1 shows that the standard deviation in test scores is much greater among schools than among districts. The variation would be even greater among individual students. As Kane and Staiger (2002a) point out, the school-level variance is only about 10-15% that of the student-level variance, a fact that has been known since the publication of the Coleman et al. (1966) report.

Class Size and Score Levels and Changes

The median grade 8 class size in the typical school nationally is roughly 70 students. This suggests that random variation may be substantial in the score that is recorded for any particular school year. Figure 1 illustrates this point, as the range in the scaled math score in grade 8 declines noticeably as the number of students enrolled in grade 8 (and taking the test) increases. This kind of heteroscedastic pattern is not uncommon in statistical analyses. Similarly, the coefficient of variation of the score levels over a 5-year period is much greater for schools with smaller classes than those with larger classes of eighth graders.

The variation in these scores is not explained solely by differences (i.e., lack of homogeneity) in the smaller schools. If small schools were otherwise homogenous, one might expect a similar range for the change in scores over time. However, this is not the case: there is also a wide range in the change in achievement scores over time among smaller schools. Figure 2 shows the average test score gain for the school years 1997-1998 to 2001-2002 in the eighth grade for math. Again, a pronounced cone-shape emerges in the range of changes as class size increases.

Schools on average scoring 25 points or more were eligible (on average) for an award in each year, assuming they also scored above 25 on average on the reading score. No school with more than 300 eighth graders scored above 25 points in average. In other words, none of the schools with more than 300 students in grade 8 on average exceeded the benchmark of 25, thus reducing significantly their incentive to participate in this accountability system.

A general conclusion from this analysis is that the variation of both the level score and the average test score gain is greater in smaller than in larger schools. To the extent that rural schools are smaller, rural areas are affected more than urban areas by this phenomenon, that is, they on average had greater odds of earning an award (and also of not earning one when they should have). See Equation 1 in Table 2, which confirms that schools in rural (or less densely-settled) areas tend to have smaller grade 8 classes than urban schools.

The next section shows that most of the change in a school's score from one year to the next is nonpersistent. That is, the change occurs for reasons that are beyond the control of schools. The smaller the school, the larger is the amount of the change over time that is explained by nonpersistent factors. This also suggests that some schools should have been rewarded when they were not (i.e., the school's change in scores would have been even smaller or more negative had the school not performed effectively).

Further Investigation into Persistent and Nonpersistent Determinants of Scores

Kane and Staiger (2002a, 2002b) propose a method for disaggregating school scores into persistent and nonpersistent (or random) components. In particular, if changes in scores from one year to the next are persistent or permanent,

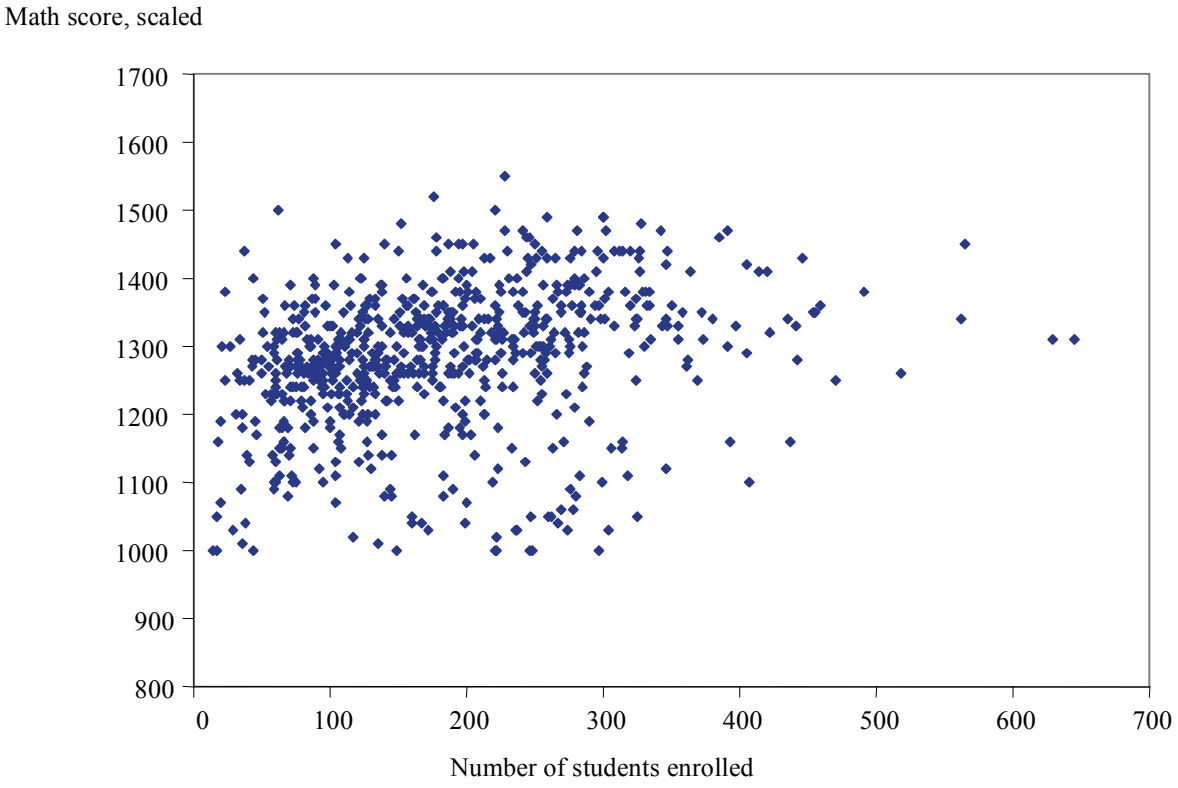

Figure 1. PSSA test score levels versus student enrollment, 1997-1998 school year, grade 8

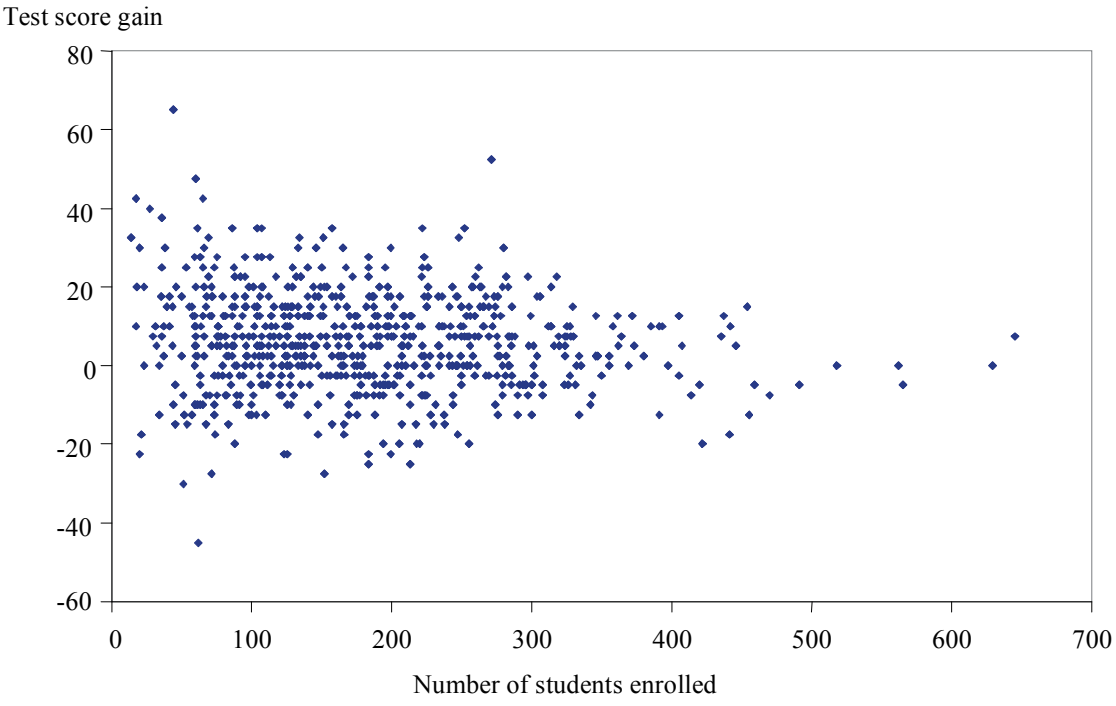

Figure 2. Average test score gain, 1997-2001 vs. student enrollment in grade 8

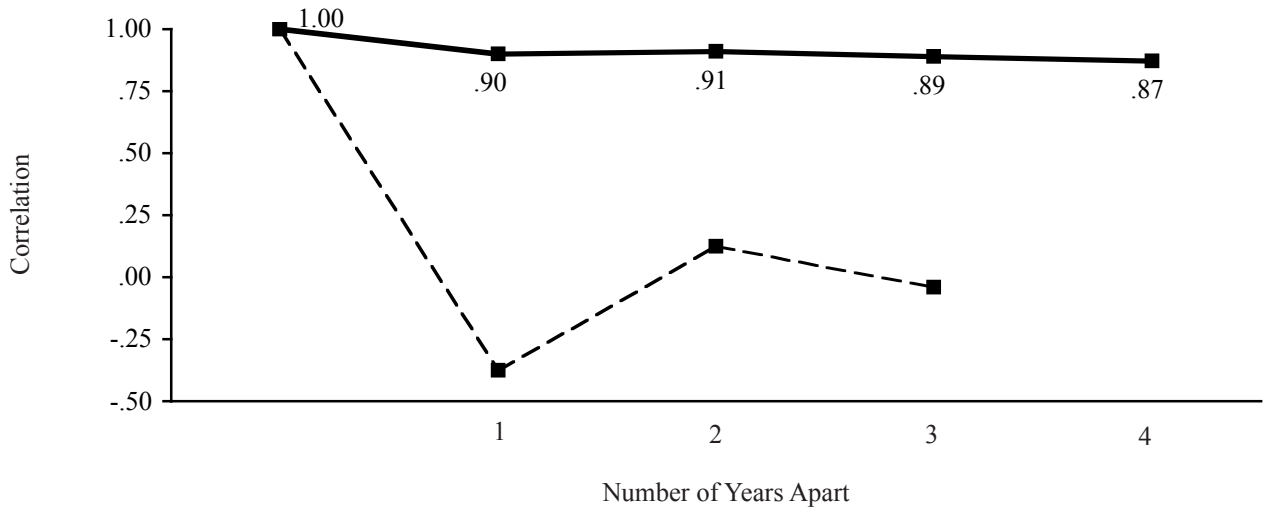

Figure 3. Test score correlations over time

then the scores will remain higher in the following year. In this case, the correlation between scores in year t and t+1 is close to 1. If, on the other hand, the change in scores is nonpersistent, the score reverts to its former value in the following year, and the correlation in changes in scores over time is -.50 (Kane & Staiger, 2001, 2002c discuss this in more detail).

In particular, doubling the correlation coefficient and multiplying by (-1) gives the share of the change in test scores explained by nonpersistent factors. The range for the coefficient is: $0 \geq \rho \geq -.50$. If none of the change in test scores is caused by nonpersistent factors (all of the change in test scores is persistent), then $\rho = 0$. On the other hand, if none of the change in test scores is attributable to persistent factors (all of the change in test scores is nonpersistent), then $\rho = -.50$. Therefore, if changes in test scores over time are a function of both persistent and nonpersistent factors, then the expected correlation among the changes in test scores from one year to the next ranges from 0 to -.50. The closer the correlation is to -.50, the larger the influence of nonpersistent (random) factors.

Figure 3 shows the correlations one and more years apart for levels of, and changes, in PSSA math scores over time. The correlation for level scores falls sharply in the first year (to .92), reflecting nonpersistent factors (by definition) and then gradually tapers off, as persistent factors slowly change (curricula, teacher turnover, etc.). Thus, schools starting out with high (low) test scores tend to end up with high (low) test scores.

In contrast, the correlation for the change scores in the first year is -.37, and it then gradually converges to zero (tapers off altogether). Three quarters of the variance in the change in math scores between one year and the next is due to temporary or nonpersistent factors (-.37/-.50 = .74). A key conclusion from this analysis is that most of the changes in the PSSA score from one year to the next are due to sampling variation and temporary factors. This is what is being rewarded under SPF in Pennsylvania.

Correlations in level and change scores, respectively, for the 50% of largest and smallest schools in Pennsylvania reveal that the relative importance of nonpersistent factors affecting test scores levels and changes over time is considerably greater among smaller schools than larger schools. In small schools, about 80% of the change in scores is nonpersistent according to these calculations, compared with 66% in the larger schools. These effects are even more pronounced when smaller units (such as quintiles or deciles) are used for the analysis.

Therefore, if there is a systematic factor within SPF or NCLB impacting rural schools—because they are on average smaller—it is that they are more likely to be rewarded for nonpersistent increases in test scores. Furthermore, just as some schools received an unjustified award, other schools should have received an award but did not because a nonpersistent factor beyond the control of superintendents happened to cause their scores to decline rather than to increase.

Poverty and PSSA Score Changes

In this section, another important factor is investigated that is typically beyond the control of superintendents, at least in the short-run. A strong and negative relationship exists between test scores and poverty rates at the level

Table 2
Regression Equations

Equation 1:	Grade8ClassSize = 338.8 + 0.840 PopDensity (18.1)[a]	R^2 = .395
Equation 2:	ScoreChange = 3.89 + 0.557 Povrate (7.49)	R^2 = .074
Equation 3:	ScoreChange = 560.8 - 0.399 BaseScore - 0.795 Povrate (15.8) (7.45)	R^2 = .265
Equation 4:	ScoreChange = 372.9 - 0.273 BaseScore - 0.177 PopDensity + 72.4 EnrollGrowth (10.7) (0.9) (2.2)	R^2 = .192

[a] *t* ratio.

of schools (with R^2 values around .62). The same picture emerges at the district level, which also shows a relationship that grew stronger statistically between 1997-1998 and 2000-2001 (this increase in the strength of the relationship did not occur at the school level, however). Thus, poverty has a definite impact on school achievement, at least in part via per pupil expenditure differences. Following education finance reform, Kentucky was able to break the strong negative correlation between poverty and student achievement (see Goetz & Debertin, 1992, 1997). A question here is why some schools with very high poverty rates are performing as well as or better than schools with lower poverty, and should these schools be rewarded differently than those with lower poverty?

However, it is also true that each school has to improve its score by 50 points to qualify for a reward, regardless of its baseline or original score. Thus, poverty should, on the surface, not be a determinant in this case. In fact, further analysis reveals that schools with higher poverty rates have statistically larger increases in scores if we do not control for initial or baseline scores (see Table 2, Equation 2). However, once we control for initial scores (in this case math), the effect of poverty on the change in scores is indeed negative and statistically significant (see Table 2, Equation 3).

Thus, schools with higher poverty levels have greater increases in scores, not holding any other variable constant. However, once we control for the initial-sample average score, the effect of poverty on changes in scores is negative. In other words, poor schools are currently catching up with wealthier schools, but only because as a group they are starting off from a lower baseline score. As these poorer schools catch up to the higher-performing schools, the effect of poverty will be to hold them back. The combination of these two factors may mean that they in practice never catch up, and that districts starting with higher scores (or poverty rates) will tend to fall behind over time.

Additional Regression Models of the Determinants of Changes in Scores

District-level models. Additional regression analysis reveals that, holding constant the baseline score and school district enrollment growth, population density had a negative but statistically insignificant effect on changes in grade 8 math scores between 1997-1998 and 2000-2001 at the district level (see Table 2, Equation 4).

Thus, rural areas had neither a greater nor smaller tendency to raise their scores over this period than urban areas. Also, districts experiencing enrollment growth (decline) experienced a statistically significant larger increase (decrease) in score changes.

Thus, a basic question is whether enrollment growth (loss) can be controlled by superintendents. If the growth or decline is due to major economic forces operating in the county, school leaders may have less control over enrollment and ought not to be punished or rewarded. However, enrollment change may be due to superintendent performance, if good schools are attracting businesses, etc., or students are leaving to join charter schools. In future research, the causes of the enrollment changes need to be sorted out. For example, enrollment changes could be modeled as a function of the creation of local charter schools, job changes, and other economic changes in the community.

Furthermore, our analysis shows that rural schools tend to have lower scores (levels) than urban schools, holding constant poverty rates, and including an interaction term, when we use population density to measure "rural." Finally, rural schools also have higher value-added scores in math between grades 5 and 8, and these differences are statistically significant, so that they are "catching up" over time. However, as noted above, they will tend to fall behind in the process of catching up, as rising scores make it more difficult for them to further improve their scores.

Table 3
OLS Results for Determinants of School-Level Test Score Gains

	Model I (abbrev.)		Model II	
	Math	Reading	Math	Reading
Constant	1.306***	1.243***	1.479***	1.512***
Grade 8 class size, 1997			0.0408*	0.0480**
Poverty rate, 1997			-0.506***	-0.703***
Grade 8 enrollment growth, 1997-2000	10.80	16.60*	8.21	11.8
Increase in poverty rate, 1997-2000	-0.423**	-0.464**	-0.839***	-1.010***
Poverty x class size interaction			0.00073	-0.0017**
Math scale score, 1997	-0.224***		-0.348***	
Reading scale score, 1997		-0.180***		-0.370***
Adjusted R^2	32.6	19.5	37.3	31.2

All parameters except for the constant scaled by 1,000.
*$p < .10$. **$p < .05$. ***$p < .01$.

School-level models. Table 3 contains models of determinants of changes in average scaled math and reading scores over the period 1997-1998 to 2000-2001 at the school level. In the first two models, changes in math and reading scores are regressed on initial scores and changes in enrollment and poverty rates. Here, we are holding constant all of the influences within schools that are captured by the initial scores, and we examine the effect of local economic changes during the time period under consideration.

Schools that experienced greater enrollment growth also had a statistically greater increase in reading scores (for math the effect was not significant statistically in this model). Higher rates of poverty had a significant depressing effect on gains, holding constant initial (1997-1998) scores. To the extent that changes in poverty status are completely beyond the control of school district superintendents, it would seem unjustified to reward or penalize them for that component of the score change that is accounted for by changes in poverty—or by changes in the effect of poverty on student performance. For enrollment growth, in the case of the reading score, this conclusion is not as straightforward, since enrollment growth may be influenced by district superintendents. As noted above, this area requires additional research.

The second set of regressions expands the regressors to include a number of pertinent interaction terms as well as other variables. Perhaps the most interesting finding is that the negative effect of poverty on changes in reading scores is exacerbated in larger schools (grade 8 class sizes). This has been documented for levels, but not for changes in scores. Larger and wealthier schools have larger improvements in reading scores. These results are sensitive to the specification. For example, when we include initial math and reading scores in the reading and math gains equations, respectively, instead of the respective gains scores, the effect of grade 8 class size is statistically different from zero in the case of the math equation as well. (For a summary of studies showing the benefits of small schools, see Lawrence et al., 2002).

Conclusion

Kane and Staiger (2002a) present three options for increasing the precision and reliability of test scores to evaluate school performance. The first is to give awards by school size category. The second is to average scores over time (at least 3 years) to reward the higher-performing schools. The MHS award option accomplishes this in Pennsylvania. A third correction is to ". . . reward schools for exceeding their expected performance, with smaller schools receiving smaller incentive payments in line with their less reliable test score measures" (p. 110).

Our descriptive and regression analysis suggests that there are presently no forces in play that systematically hurt rural or smaller schools in terms of NCLB in general and Pennsylvania's SPF in particular. In fact, smaller schools have an advantage over larger schools in securing an award in any given year, because their scores bounce around more than is true for larger schools. It may also be the case, however, that small or rural schools can more efficiently make improvements in response to incentives. In particular, a smaller district may be able to more easily make district-wide curricular changes than larger districts. This needs to be sorted out in future research.

To the extent that rural schools are also poorer, and they have lower baseline scores, it may be easier for rural schools to raise their scores, at least initially. We can expect this to change once they catch up with the other schools, however. Further, in the smaller schools, 80% of the increase in scores is estimated to be caused by nonpersistent factors, that is, factors that generally cannot be controlled by school staff and administrators. Likewise, some schools likely should have received an award in a given year, but failed to get one because nonpersistent factors pushed their scores down to a level that disqualified them.

Finally, our regression analysis reveals that poverty significantly depresses the gains achieved by schools, once we control for initial scores. At the same time, small schools are able to counteract the negative effect that poverty exerts on changes in scores over time. Also, increases in poverty in a district, a phenomenon that would seem to be completely beyond the control of superintendents, also disadvantages schools in terms of gains in scores.

References

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Goetz, S. J., & Debertin, D. L. (1991). A caution about using school outputs in educational production functions [Anthology]. *Atlantic Economic Journal, 19*, 62.

Goetz, S. J., & Debertin, D. L. (1992). Rural areas and educational reform in Kentucky: An early assessment of revenue equalization. *Journal of Education Finance, 18*, 163-79.

Goetz, S. J., & Debertin, D. L. (1997). Local economic conditions and KERA (the Kentucky Education Reform Act). In *Center for Business & Economic Research 1996 annual report* (pp. 25-30). Lexington, KY: University of Kentucky, Department of Economics.

Kane, T. J., & Staiger, D. O. (2001). *Improving School Accountability Measures* (NBER Working Paper No. 8156). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2002a). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91-114.

Kane, T. J., & Staiger, D. O. (2002b). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on educational policy* (pp. 235-283).Washington, DC: Brookings Institution.

Kane, T. J., & Staiger, D. O. (2002c, June). Racial subgroup rules in school accountability systems. Paper presented at the conference Taking Account of Accountability: Assessing Policies and Policy, Harvard University, Cambridge, MA.

Lawrence, B. K., Bingler, S., Diamond, B. M., Hill, B., Hoffman, J. L., Howley, C. B., Mitchell, S., Rudolph, D., & Washor, E. (2002). *Dollars and sense: The cost effectiveness of small schools, knowledge works foundation.* Cincinnati, OH: Knowledge Works Foundation.

Pennsylvania Department of Education. (2000). *1999-2000 Pennsylvania school performance funding program.* Retrieved from http://www.state.pa.us

Pennsylvania Department of Education. (2002). *School performance funding school-by-school results.* Retrieved from http://www.pde.state.pa.us/k12_initiatives/cwp/view.asp?a=173&Q=85753&PM=1

.