School Grades Based on Standardized Test Scores: Are They Fair?

Harriet A. Stranahan
University of North Florida
Jacksonville, FL 32224
hstranah@unf.edu

J. Rody Borg
Jacksonville University
Jacksonville, FL 32211
rborg@ju.edu

Mary O. Borg
University of North Florida
Jacksonville, FL 32224
mborg@unf.edu

Abstract

In 1997, Florida passed legislation that created the School Recognition Program, which gives schools letter grades from “A” through “F.” Given the important impact of these school grades, it is essential to know if a school’s grade depends more on the intrinsic qualities of the school or on the qualities of the individual students who go to that school. Using a sample of 15,100 elementary students, our study is an attempt to determine this.

Introduction

There is a large multi-disciplinary literature that exists on the factors that affect academic achievement and the closely related topic of school effectiveness. This literature began in 1966 with the controversial Coleman report that implied that school inputs have almost no effect on schooling outcomes. According to the Coleman Report, family background is clearly the most important and dominant predictor of educational attainment. Eric Hanushek (1986,1989, 1994) has been the leading proponent of the view that increased spending on school resources has little, if any, substantive pay-offs in terms of student achievement. Other education researchers strongly disagree. Grissmer, Flanagan, Kawata and Williamson (2000) cite evidence that indicates measurement errors are the primary reason that Hanushek and his followers obtain their results (Rothstein and Miles, 1995; Ladd, 1996.) Furthermore, even though increases in overall educational spending may appear to have no effect on overall educational achievement, as measured by average student test scores, there are some specific categories of spending that show dramatic effects on the test scores of certain groups of students. For example, when increases in educational spending are used to reduce class size for minority and economically disadvantaged students, the experimental evidence suggests that the test scores of this group of students improve (Finn and Achilles, 1999; Krueger, 1999).

The controversy over the degree to which educational resources can make a difference in the educational attainment of students has taken on new urgency today as states grapple with educational reform that relies heavily on standardized test scores as a measure of school effectiveness. In 1997, Florida passed legislation that created the School Recognition Program, which gives schools letter grades from “A” through “F.” The primary criteria by which these letter grades are assigned are individual students’ scores on the standardized Florida Comprehensive Assessment Test (FCAT) in a given year. In 2002, the formula for determining a school’s grade was modified so that changes in the school’s FCAT scores from the previous year were also factored into a school’s grade. Other data such as the percentage of students tested, attendance and discipline data, and dropout rates also affect a school’s grade. However, the major determinant of a school’s grade still depends on the FCAT scores of the students who attend that school during the current academic year. The grade that a school receives has a major financial impact on the funding of public schools. Schools that receive “A” grades receive $100 more per full-time equivalent (FTE) student than other schools, and schools that receive two F grades within a four-year period may see their funding fall precipitously because their students will be eligible to transfer to other schools.

Given the important impact of these school grades, it is essential to know if a school’s grade depends more on the intrinsic qualities of the school or on the qualities of the individual students who go to that school. If the school itself is contributing to the high or low performance of the students on the FCAT, then the economic incentives are justified. On the other hand, if the individual qualities of the students, themselves, are the main determinant of the students’ test scores, as Hanushek and his followers believe, then rewarding or penalizing the school is not justified since schools have no control over the student population that they serve. Furthermore, reducing the public resources that go to the lowest performing schools, which serve a disproportionately large number of minority and economically disadvantaged students, may have serious negative consequences to the test scores of these students, since recent experimental studies (Finn and Achilles, 1999; Krueger, 1999; Nye, Hedges, and Konstantopoulos 1999; Finn, Gerber, et. al., 2001) show that increased resources in these schools significantly improve their performance on standardized tests. Therefore, it is imperative that we understand the factors that affect student FCAT scores and the letter grades that schools receive.

Data

The Duval County (Jacksonville, FL) public school administration has provided us with Florida Comprehensive Assessment Test (FCAT) scores for 4th and 5th grade students who took the test in the 1999-2000 school year. These data also include demographic information on race, gender, number of times the student has withdrawn from school (an indicator of student mobility), as well as the gifted status and free or reduced price lunch eligibility of each student. We supplement this individual student demographic data by using the student's address to link each student with census block level demographic data. This allows us to create a demographic profile for each student using the census block level values for variables such as parents' education levels. In addition, the Duval County school system collects a variety of school-level data such as student absences, number of teachers with advanced degrees, teachers' years of experience, proportion of teachers newly hired, magnet school indicators, proportion of students in the school who receive free or reduced price lunch and proportion of teachers and staff who rate the principal as highly effective and a strong leader. These data allow us to specify a number of school factors that may affect student performance on the FCAT. Therefore, we have available a wide variety of family background and school specific factors that have not been previously included in one model. The mean values of the variables are shown in Table 1. The variable means are given for the whole sample, as well as for the A-rated and the non-A-rated schools.

Table 1. Descriptive Statistics Duval County Elementary School Students

Student Characteristics:	MEANS		
Educational Attainment of Adults in Household:	All Schools	Non-A-Rated	A-Rated
Less Than 9th Grade	0.068	0.0729	0.047
9th Grade to 11th Grade	0.155	0.1654	0.1131
High School Graduate	0.315	0.3246	0.2756
African American	0.241	0.258	0.1684
Hispanic	0.030	0.032	0.023
Student is Eligible for Free/Reduced price lunch	0.444	0.4849	0.2685
Number of Student Withdrawals from School	0.524	0.526	0.5174
Male	0.488	0.4908	0.4749
Student in Gifted Program	0.059	0.0458	0.1151
% Students Eligible for Free/ Reduced Price Lunch	0.5318	0.5701	0.3661
% Students Absent More than 21 Days	0.075	0.0788	0.0581
School Characteristics:			
Magnet School	0.5097	0.4778	0.6474
Average Class Size	24.200	24.1535	24.3791
% of Teachers with Advanced Degrees	0.302	0.2962	0.3251
Teachers Average Years of Experience	14.33	14.2584	14.6269
% of Teachers who are Newly Hired	0.114	0.1194	0.0879
% Teachers Rating Principal A or B	0.775	0.7728	0.7822
Percentile Norm Referenced Reading FCAT	54.41	52.64	62.19
Percentile Norm Referenced Math FCAT	58.99	57.16	66.95
Number of Students in the Sample	15161	12311	2850

Statistical Analysis

We first estimate regression equations explaining student reading and math FCAT scores for the entire sample of 4th and 5th graders in Duval County. The results for these regressions are shown in columns 1 and 2 of Table 2. The dependent variable is percentile scores for Duval County students on the reading and math portions of the FCAT standardized tests given to 4th and 5th graders. This portion is nationally normed; therefore, we are estimating the student's percentile rank on the FCAT relative to 4th and 5th grade students nation-wide.

Table 2. The Estimated Models for Standardized Math and Reading Test Scores

Independent Variable	All Schools Math (1)	All Schools Reading (2)	"A" Schools Math (3)	"A" Schools Reading (4)	Non "A" Math (5)	Non "A" Reading (6)
Constant	68.014***	80.404***	653.25***	689.95***	64.674***	75.796***
Less than 9th Grade	-20.204***	-17.506***	-35.912***	-33.802***	-17.338***	-15.497***
9th – 11th Grade	-8.620***	-11.404***	-9.832	-5.4295	-9.065***	-13.201***
High School Graduate	-12.925***	-9.816***	-3.366	3.025	-13.681***	-11.439***
African American	-14.057***	-12.834***	-13.966***	-14.778***	-14.05***	-12.358***
Hispanic	-4.546***	-5.159***	-3.239	-4.789*	-4.685***	-5.149***
Free Lunch Eligible	-6.618***	-6.716***	-7.166***	-5.87***	-6.427***	-6.800***
Number of School Withdrawals	-0.321	-4.439***	1.947	-8.112**	-0.743	-3.8233***
Male	0.38131	-4.305***	-0.366	-4.184***	0.536	-4.361***
Gifted	23.524***	26.697***	18.156***	21.031***	25.342***	28.554***
% in School Free Lunch Eligible	-11.387***	-13.421***	-23.669***	-33.7***	-11.125***	-12.455***
Magnet School	0.358	1.054***	3.746*	7.040***	-0.136	0.255
Average Class Size in School	0.873***	-0.2344	-43.582***	-47.098***	0.998***	0.164
Average Class Size Squared	-.0015***	.0002	0.925***	0.989***	-.00170***	-.0003
% Teachers with Advanced Degrees	-28.11***	-22.428***	-229.09***	-180.32***	-32.176***	-36.436***
Average Years of Teacher Experience	-0.735***	-0.611***	-6.287***	-5.897***	-0.742***	-0.788***
Interaction Advanced Degrees*Years of Experience	1.883***	1.501***	16.936***	14.333***	2.161***	2.392***
% of Newly Hired Teachers	-14.205***	-6.598***	-54.714*	-22.498	-12.717***	-5.666***
% Principal Leadership A or B	29.795***	28.520***	117.56***	105.57***	32.647***	31.56***
% Principal Leadership A or B Squared	-18.853***	-19.728***	-102.29***	-95.276***	-20.744***	-21.714***
Number of Observations Adjusted R2	15161 0.273	15161 0.276	2850 0.261	2850 0.277	12311 0.260	12311 0.264

*** Indicates the coefficient P-value <.01, ** coefficient .01< P-value <.05 and * .05< P-value <.10

As expected, a key indicator of student success on both the reading and math portions of the FCAT is parental education level. Specifically, students whose parents have not completed high school (about 20% of the sample) have significantly lower test scores across the board. Further, children whose parents are college educated have significantly higher scores than those whose parents have a high school diploma only. These results are consistent with other studies showing that the intergenerational transmission of human capital is a very important component of students' readiness and aptitude for learning (Datcher, 1982; Hill and Duncan, 1987; Krein and Beller, 1988; Case and Katz, 1991; Duncan, 1994; Graham, Beller and Hernandez, 1994; Haveman and Wolfe, 1995).

Along with education, parental income is also an important predictor of test scores. The variable indicating a student's eligibility for a free or reduced price lunch is a proxy for poverty status, or at the very least, low household income. As expected, individual students who are eligible to receive free or reduced price lunches perform significantly worse on both math and reading FCATs in all regressions estimated. In addition to the variable indicating the free or reduced price lunch status of each student in the sample, we include a variable indicating the percentage of students within each school who receive free or reduced price lunches. This variable serves as an important indicator of the environment of each school. The negative and significant coefficient on this variable suggests that students of any background who attend schools with high proportions of low-income students tend to have lower FCAT test scores.

In addition to parental education and income variables, race and gender are individual student characteristics that significantly affect FCAT scores, for the sample as a whole. Male students score significantly lower on FCAT reading tests, although gender is not a significant indicator of FCAT math performance for 4th and 5th graders. African American and Hispanic students have significantly lower FCAT scores in both math and reading than Caucasians and students of other races.

The student's mobility rate is measured by the number of times the student withdrew from school. Our sample contains only students who were at the same school in both the beginning and end of the school year. Therefore the mobility rate in our sample indicates that a student withdrew and was readmitted one or more times during the course of the school year. Number of withdrawals had the expected negative and significant effect on FCAT reading scores, but surprisingly it did not have a significant effect on FCAT math scores for the 4th and 5th graders.

We also include a variable indicating whether the student was placed in the gifted program. Students enrolled in the gifted program have greater ability as measured by standard intelligence tests. As expected, the results show that students in the gifted program have significantly higher math and reading FCAT scores than students who are not in the gifted program.

There are a number of school factors that are important predictors of student FCAT scores, as well. Originally designed as a way to integrate and diversify the student body of predominantly minority schools, about 50% of the students in Duval County attend a school with a designated magnet program. Students from any part of the county can go to these schools if they apply to attend and are admitted into the magnet program. Magnet schools often attract highly motivated students because they offer special instruction for gifted and talented students as well as special programs such as Montessori education, language immersion, science, technology and special instruction in the arts. Not all students that attend these schools participate in the magnet program, many are neighborhood children enrolled in regular education. The regression results suggest that students who attend a magnet school have higher FCAT reading scores, but math scores at these schools are not significantly different than the math scores of students in non-magnet schools.

Our results also show that teacher characteristics are important predictors of student success. Students attending schools with higher rates of teacher turnover (a higher proportion of teachers new to the school) have lower FCAT math and reading scores, all else equal. High turnover is commonly associated with difficult teaching conditions or poor school management. Teacher education levels and years of experience also affect FCAT scores. Teachers at Duval County schools have an average of 14 years of teaching experience and more than 30% of the teachers have earned a Masters, Ph.D. or other post-graduate specialist degree. We include an interaction term for these two variables in order to find out, for example, whether having a more educated teaching staff might change the effect of years of teacher experience on student success in reading and math. The interaction is highly significant in each of the regression models. The coefficients on education and experience suggest that these two factors are complements and work together to increase students' FCAT scores. Specifically, students are most successful on math and reading FCAT's when their school faculty has above average years of experience and above average education levels. A more educated faculty tends to increase the effectiveness of experience (and vice versa). Schools with below average teacher experience and education, not surprisingly, have students that perform more poorly on both reading and math FCATs. [1]

The significant coefficients on the class size and the square of class size variables suggest that the relationship between class size and FCAT math scores is curvilinear. Evaluating these coefficients at the mean or any reasonable value of class size leads to the result that students in schools with larger class sizes do better on the FCAT. Clearly, this finding runs counter to the common perception that students learn better in smaller classes. We believe that the positive relationship in our regression reflects the fact that average class size is correlated with school quality for the aggregated sample. In Duval County, class size rarely gets bigger than 32 students per class, and the biggest classes tend to be in the newer schools in the fast growing suburban neighborhoods of Duval County. Class sizes are smallest in the older inner-city schools of the county where student population is declining. For example, in the 1999-2000 academic year, Duval County schools that received school grades of C, D and F had average class sizes of 24, 21 and 19, respectively. In other counties across the state as well, failing schools located in the inner cities typically have smaller average class sizes. For the sample as a whole then, the coefficient on class size is positive. However, this result changes later when we run regressions on A-rated schools only. Interestingly, FCAT reading scores were not affected by average class size.

Schools with strong leadership, indicated by a high percentage of teachers giving the principal an A or B rating, had significantly higher FCAT math and reading scores. We also include a squared form of the leadership variable in the regression equation. Its significance indicates that the relationship between the principle's leadership and FCAT scores is curvilinear. When we evaluate the coefficient at the means of the data or any reasonable value, we find a positive and significant relationship between principal leadership and student performance on the FCAT.

"A" Schools versus Non "A" Schools

In an effort to discern differences in how school and student related factors affect student performance in A versus non-A schools, we estimate two separate regressions explaining 4th and 5th grade math and reading FCAT scores in A-rated versus non-A rated schools. The results of these regressions are shown in columns 3-6 of Table 2. Not surprising, because school grades depend largely upon FCAT scores, schools attaining As have significantly higher scores compared to those ranking in the non-A categories. Table 1 shows that the sample average scores on the norm referenced math and reading FCATs jump by 10 percentage points for A versus non-A schools. This means that reading FCAT scores are 18% higher and math FCAT scores are 17% higher in A schools than in non-A schools, on average.

Many of the same factors that impact student FCAT scores for the entire sample are also significant when the sample is separated into A versus non-A schools. We will discuss only the variables that have differential effects in the A versus non-A schools since the results of the total sample have already been discussed.

Among A-rated schools, students who attend a magnet school score significantly higher on math and reading FCATs than the students in A-rated schools without magnet programs. This is an interesting result because every student at a magnet school is not a highly motivated student who is bused in from another part of the county. Many of the students attending a magnet school are kids who just happen to live in the neighborhood where the school is located, and most magnet schools are located in disadvantaged neighborhoods. Another interesting result is that students who attend magnet programs in non-A rated schools do not score significantly better on the FCAT than the students in the other non-A schools. This indicates that magnet schools have the potential "to lift all boats" only if they are highly effective. In other words, magnet schools can improve the test scores of all of their students if they are good enough to boost their school's rating to an A.

Our results regarding class size in the A versus non-A schools are also interesting. As mentioned previously, in the model that includes all schools, we find that class size has a positive and significant effect on FCAT scores, indicating that FCAT scores are higher when class sizes are large. However, when we estimate models for the A and non-A schools separately, we find this perverse effect only in the non-A schools. This reflects the fact that inner city schools have smaller enrollments and smaller average class sizes, and they tend to be the schools with the lowest FCAT scores. On the other hand, in the well-attended, sometimes over-crowded, A-rated schools, class size (fit with class size squared) has the expected negative and significant effect on FCAT math and reading scores. Students who are in the smaller classes within the subset of A schools do have higher FCAT math and reading scores, on average. This suggests that reducing class size does in fact lead to positive educational outcomes when the spurious correlation between small class sizes and failing inner city schools is removed from the equation.

Students in non-A schools with higher rates of teacher turnover (a higher percentage of newly hired teachers) have lower reading and math FCAT scores. Teacher turnover in the A-rated schools negatively impacts FCAT math scores but not FCAT reading scores. These results indicate that teacher turnover has more serious effects in the non-A schools, and unfortunately, the non-A schools also experience more teacher turnover than the A schools.

Marginal Effects in A versus Non-A Schools

A clear pattern emerges from these results -- both school characteristics and student characteristics are important predictors of student success on the FCAT across all types of schools. However, looking at these regressions when they are separated into A versus non-A schools, some of the variables have very similar marginal effects in both A and non-A schools, and other variables produce very different marginal effects in the two types of schools. By evaluating the same variables in the A and non-A regression equations at their mean values (or at 0 and 1, for the dummy variables), we can identify whether the same variable has a similar or different marginal effect in each type of school.

In general, most of the variables that reflect the demographic characteristics of the students seem to translate into similar differences in FCAT scores in both A and non-A schools. For example, whether free lunch-eligible students go to an A school or a non-A school, their FCAT math and reading scores are going to be about 6-7 percentage points lower, on average, than an identical student who is not eligible for the free lunch program. Likewise, whether attending the A or the non-A rated schools, African Americans' scores are about 12 percentage points lower, gifted students about 20 or so points higher, and male students' reading scores are about 4 points lower than female scores. Among the school related characteristics, the marginal effect for years of teacher experience is small but very close in magnitude in A and non-A schools (0.55 and 0.48 evaluated at 40% of teachers with advanced degrees).

These results suggest that gifted, African American, male, or free lunch-eligible students can expect the same marginal differences between their test scores and the scores of other students who attend the same type of school, whether they are in A or non-A-rated schools. Likewise, years of teacher experience is equally effective in A and non-A schools. It is important to note, however, that the marginal effects of these variables are similar in the two types of schools, but the levels are not. When we say that a free-lunch eligible student can expect to earn 6-7 percentage points less than the other students who are not eligible for free lunch, the other students in an A school have scores roughly 10 percentage points higher than the other students in a non-A school.

Whereas some variable coefficients are remarkably similar, other coefficients are quite different between the A and non-A school regressions. As already noted, the effect of class size is substantively different in A versus non-A schools. The results suggest that smaller class sizes will not improve FCAT scores in non-A schools, but would have a positive effect on performance in A-schools. We should qualify this result to say that these results are only valid within this sample's range. Very small class sizes, beyond the scope of our sample, may result in different outcomes.

There is also a significant difference in the effect that advanced degrees for teachers has on FCAT scores between the A and non-A schools. Evaluating these coefficients close to the mean in the FCAT math regression (at 15 years of experience), the marginal effects are 24.9 and 0.24 for A versus non-A schools, respectively.[2] This suggests that advanced teacher degrees do translate into significant learning gains for A schools, but have little or no effect on student performance in non-A schools.

There is also a significant difference in the effectiveness of principal leadership. Evaluated at the means of the data, the marginal effects on principal leadership are 38.7 and 16.6 in A versus non-A schools, respectively. Strong principal leadership is an important component for school success in A-rated schools but is significantly less effective in non-A schools.

In summary, our results show that A-schools are not more effective in teaching minority students from low income or less educated households, as evidenced by the similar marginal effects of the student demographic variables in the regression models of the A and non-A schools. However, one key benefit of A-schools appears to be that they possess an environment that can better translate traditional inputs such as small class sizes, teacher education and strong principal leadership into effective student learning. This result begs the next question. Why are these schools more effective at utilizing their inputs? Are the A schools really doing a better job in the classroom, or is the source of their success a more ready-to-learn student body?

A Schools with Non-A Students and Vice Versa: An Oaxaca Decomposition

To answer these questions, we will use our regression results in an Oaxaca decomposition (Oaxaca, 1973). This technique uses the estimated regression equation for the A level schools to predict the average percentile FCAT scores in reading and math for a typical student in a non-A school. This provides additional insight into whether these schools' A-ratings are due to a superior mix of school attributes or to the characteristics of their student body.

The values of all school characteristics (such as class size, teacher attributes, etc.) in the estimated regression equation are the average values for the sample of A schools, but the values of all student characteristics (such as race, free or reduced price lunch status, parental education levels, etc.) in the estimated regression equation are the average values for the sample of non-A schools. Table 1 shows that the non-A schools tend to have lower parental educational attainment, fewer gifted students, and a higher percentage of students receiving free lunches.

The results of the Oaxaca decomposition are summarized in Table 3. We find that A schools with the student characteristics of the non-A schools have much lower predicted test scores. Predicted math and reading percentile scores dropped to 56.2 and 49.9, respectively. These scores are slightly lower (although not statistically different) than the actual average FCAT scores earned by the non-A schools (57.1 and 52.6, respectively). This suggests that A schools are no more effective than non-A schools at teaching this more disadvantaged student population.

Table 3: Results of the Oaxaca Decomposition Analyses

Type of School with Type of Student Population	FCAT--Math Predicted Test Scores (National Percentiles)	FCAT--Reading Predicted Test Scores (National Percentiles)
A-Rated School with its Own Student Population	66.92	62.17
A-Rated School Teaching an Average Student From Non-A Schools	56.25	49.90
Non-A School with its Own Student Population	57.14	52.63
Non-A School Teaching the Average Student from an A-School	65.57	61.61

If we do the same sort of Oaxaca decomposition for non-A rated schools, giving the non-A schools the average student characteristics of the A rated schools, the predicted test scores for the non-A schools rise dramatically. Math and reading FCAT scores jump to 65.6 and 61.6, respectively. These predicted values reinforce the results of our previous analysis of the magnitudes of the regression coefficients. Both suggest that A and non-A schools do an equally effective job with socioeconomically advantaged student populations and an equally ineffective job with disadvantaged student populations. Even if the school environment in an A-school could be maintained with the non-A school student population (as assumed in the Oaxaca decomposition), A-rated schools do no better at educating disadvantaged students. They have virtually the same predicted math and reading FCAT scores as the non-A schools have with the same population of students.

The Probability That a School Will Earn an A Rating

Our regression analysis and predicted values from the Oaxaca decomposition analyses clearly show that both the A schools and the non-A schools produce almost identical FCAT scores given similar student characteristics. We also found that the usual inputs associated with positive educational outcomes including advanced teacher education, strong principal leadership and small class sizes are more effective in A-schools; in other words, these same inputs just don’t yield the same results in non-A schools. With millions in funding at stake in the School Recognition Program, the important issue is whether student or school characteristics are the key to obtaining an A rating. What factors affect the probability of getting an A-rating and the subsequent much-needed school funding?

To help answer this question, we estimate the probability that a school with a particular set of student and school characteristics will earn an A. To do this, we use a probit estimation technique because our dependent variable (school earns an A or school does not earn an A) is dichotomous. The results of the probit regression are shown in Table 4.

The results show that both school and student body characteristics are strong predictors of school grades. The impact of each of the variables is similar and largely consistent with the results reported in the regressions in Table 2. Students who come from households with less educated parents are less likely to attend an A-rated school. Not surprisingly, schools that have a higher percentage of their student population receiving free or reduced price lunches, that is students close to the poverty level, are less likely to receive an A rating.   Schools that have more students with high rates of absenteeism are also less likely to receive an A.   As expected, students in gifted programs are more likely to attend an A-rated school.   All else equal, schools with a magnet program are more likely to receive an A rating, which may help explain why African Americans are more likely than other races to attend an A-rated school. Schools with relatively larger class sizes and lower teacher turnover are also more likely to receive an A grade. This is explained by the fact that both of these characteristics – large class size and low teacher turnover rates – are more likely to be found in suburban schools than in inner-city schools.

Table 4. The Probability of Attending an A-rated School

Independent Variable (1)	Probit Coefficients (2)
Constant	-0.933***
Less than 9th Grade	-0.111***
9th – 11th Grade	-0.002
High School Graduate	-0.165***
African American	0.002***
Hispanic	-0.0007
Gifted	0.002***
% In School Absent More than 21 Days	-0.0005*
% In School Receiving Free Lunch	-0.238***
Magnet School	0.003***
Average Class Size in School	0.129***
Average Class Size Squared	-0.0003***
% Teachers with Advanced Degrees	-0.706***
Teachers Average Years of Experience	-0.002***
Interaction Years of Experience * Advanced Degrees	0.005***
% Teachers who are Newly Hired	-0.212***
% Teachers Rating Principal Leadership as A or B	-0.265***
% Principal Leadership Squared	0.189***
Number of Observations	15161

*** Indicates the coefficient P-value <.01, ** coefficient .01< P-value <.05 and *.05< P-value <.10

The coefficients on teachers' average years of experience and percent of teachers with advanced degrees show results similar to those in the FCAT math and reading regressions.[3] Schools whose teachers have more than the average number of years of experience are more likely to receive an A rating. Similarly, schools who employ more than the average number of teachers with advanced degrees are also more likely to get an A grade. Years of experience and advanced degrees are complements, each increasing the impact of the other on the probability of receiving an A rating.[3] Strong principal leadership is the only variable whose sign is somewhat inconsistent with the regression results reported in Table 2. The probit results suggest that schools with the most highly rated principals are less likely to receive a grade of A. Strong leadership alone is not enough to earn the school an A-rating; indeed, it may be a prerequisite for managing schools with the more difficult to teach students.

Table 5 shows the predictions from the probit model using a school's own student population as well as the student population of the other type of school. The question we wish to explore is whether the schools that receive A grades would be equally likely to get the same grade (and state funding) if they had to educate students like those attending the non-A-rated schools. The probability of earning an A rating drops more than three-fold from a 47% chance of earning an A to a 12% chance of earning an A, if these schools were teaching the typical student found in the non-A schools. Conversely, the probability that non-A-rated schools would earn an A rises more than four-fold, from 9% to 40% when non-A schools are given the A school's typical student body.

These results suggest that the state is rewarding student attributes rather than school attributes. Denying additional state funds, as the School Recognition Program has done, to schools with students who come from households with low income and low levels of parental education may actually be reinforcing a system of failure rather than providing a remedy to these failing schools.

Table 5. Predictions Using the Probit Model

	Probability of Achieving an A-Rating
A-Rated School with its Own Student Body	0.47
A-Rated School Teaching the Average Students From Non-A Schools	0.12
Non-A School with its Own Student Body	0.09
Non-A School Teaching the Average Students from an A-School	0.40

Summary, Conclusion, and Policy Recommendations

We find that most student and school variables have remarkably similar impacts on test scores in A and non-A schools, while a few variables have substantially different effects in the two types of schools. For example, the results suggest that smaller class sizes will not improve FCAT scores in non-A schools, but would be a very effective strategy for raising scores in A-schools. Likewise, advanced teacher degrees do translate into significant learning gains for A schools, but have little or no effect on student performance in non-A schools.

These results also shed light on the debate about whether school resources impact student learning. The key benefit of A-rated schools appears to be that they provide an environment that can better translate traditional inputs such as small class sizes, teacher education and strong principal leadership into effective student learning. We find that traditional school resources have less of a positive impact in schools serving students from disadvantaged households. In the non-A schools, the standard educational model emphasizing class size teacher education, teacher experience and strong principal leadership may be helpful, but it is not enough to raise these students' learning outcomes, as measured by standardized test scores. In the non-A schools, spending money on new educational strategies that deal specifically with the problems faced by children living in poverty may be more effective than spending additional money on traditional school inputs.

So what is the state rewarding with the School Recognition Program -- schools with more effective teachers and administrations or schools that are lucky enough to teach a population of students coming from higher socioeconomic households? Unfortunately, our results suggest the latter answer. Our Oaxaca decomposition analysis shows that both the A schools and the non-A schools produce virtually identical FCAT scores when they teach the same population of students. Our probit analysis confirms that the A-rated schools would find their probabilities of earning an A fall by more than 75% if they had student populations that matched those of the non-A rated schools. On the flip side, the chance of earning an A in a non-A-rated school would increase by 344%, if given students like those who attend the A-rated schools.

All evidence points to the fact that Florida's funding mechanism is rewarding schools based upon the composition of their student bodies. A funding formula that rewards the higher socioeconomic status of the student body in otherwise equally effective schools is inequitable (it is a regressive subsidy program). It is also inefficient because it provides an incentive for teachers and other school resources to flow to the schools that need help the least. The additional funding provided to many A-rated schools has often been used to fund teacher bonuses. When the best teachers, principals and staff are drawn to the A-rated schools, disadvantaged schools may be left worse off in the long run.

The policy recommendation that results from this analysis is clear. Schools should be graded on how well they teach their own populations, not on how their population of students compares to others. This can be accomplished by measuring a school's effectiveness by the amount of improvement that occurs in the same set of students over the course of a year. The change in each individual student's test scores should be calculated as the student advances from grade to grade. If the student has remained in the same school for two consecutive years, then the change in that student's test score is a valid measure of the school's effectiveness, not the FCAT level.

Since the 2001-2002 school year, the criteria for grading schools in Florida has been amended to incorporate changes in aggregate student test scores, but schools with high FCAT levels across the board still have a strong advantage in receiving a high grade. The new school grades are determined by a point system as follows: one point is awarded for each percentage of the school's students who score at or above grade level on the FCAT reading, math, science and writing tests, one point is awarded for each percentage of students making learning gains in reading and math, and one point is awarded for each percentage of the lowest performing students who make learning gains in math and reading. The state then determines what the point range for each grade will be. The letter grade is assigned based on the total number of points earned by each school, but other criteria such as the percentage of students tested, attendance and discipline data can reduce a school's grade. Schools with high performing students can still get an A-rating as long as their students' scores remain constant, and a significant percentage of their students do not digress. Schools serving disadvantaged populations are still penalized for having low FCAT scores, but they are given some credit for year to year improvement in the scores of their lowest performing groups. However, it is important to note that school grades are still based on the performance of the students who happen to be attending the school on the day the FCAT is given. Individual student progress is not tracked for purposes of assigning school grades. Because student turnover rates are much higher at schools with lower socioeconomic status populations, this presents another disadvantage to the lower performing schools.

The implications of Florida's School Recognition Program have importance for the entire United States because the Federal No Child Left Behind Act measures school performance in essentially the same way. Certainly school accountability is a worthy goal, but if schools are going to be graded, they should be graded in a way that measures how much improvement they provide in individual students' test scores rather than on the aggregate level of their students' test scores in a particular school year. Furthermore, any reward system for schools based on school grades should result in an equitable distribution of resources. Only then can we say that no child is left behind.

References

Case, A., & Katz, L. (1991). The company you keep: The effects of family and neighborhood on disadvantaged youths. *National Bureau of Economic Research Working Paper 3705.*

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

Datcher, L. (1982). Effects of community and family background on achievement. *Review of Economics and Statistics,* 64,1, 32-41.

Duncan, G. J. (1994). Families and neighbors as sources of disadvantage in the schooling decisions of white and black adolescents. *American Journal of Education,* 103, 20-53.

Finn, J. D. &. Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, and misconceptions. *Educational Evaluation and Policy Analysis*, 20, 97-109.

Finn, J.., S. Gerber, C. Achilles, & J. Boyd-Zaharias (2001). The enduring effects of small classes. *The Teachers College Record*, 103, 2, 145-183.

Graham, J,, Beller, A. H., & Hernandez P. (1994). The effects of child support on educational attainment. In I. Garfinkel, S. McLanahan, & P. Robins. (Eds.), *Child support and child well-being* (pp. 317-54). Washington, D.C.: Urban Institute Press.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us.* Santa Monica, CA: RAND.

Haveman, R. & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature,* 33, 32-41.

Hanushek, E. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24, 1141-1176.

Hanushek, E. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, 18, 454-51.

Hanushek, E. (1994). *Making schools work: Improving performance and controlling costs.* Washington, DC: Brookings Institution.

Hill, M., & Duncan G. I. (1987). Parental family income and the socio-economic attainment of children. *Social Science Research,* 16, 39-73.

Krein, S. F., & Beller, A. H. (1988). Educational attainment of children from single-parent families: Differences by exposure, gender and race. *Demography,* 25, 221-34.

Krueger, A. B. (1999). Experimental estimates of educational production functions. *Quarterly Journal of Economics*, 104, 497-532.

Ladd, H. F. (1996). *Holding schools accountable*. Washington, DC: Brookings Institution.

Nye, B., L. Hedges, & S. Konstantopoulos. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis*, 21, 2, 127-142.

Oaxaca, R. (1973). Male and female wage differentials in urban labor markets. *International Economic Review,* 14, 693-709.

Rothstein, R., & Miles, K. H. (1995). *Where's the money gone? Changes in the level and composition of education spending*. Washington, DC: Economic Policy Institute.

Footnotes

[1] When we include an interaction term in the regression such as (% Teachers with Advanced Degrees (Adv Deg) * Teachers Average Years of Experience (YrsExp)), we can never interpret the effect of AdvDeg (or YrsExp) without also simultaneously considering the interaction term. It is incorrect to look at the variable's 'individual impact'. The significant interaction coefficient implies that these two variables act in concert to affect Reading and Math Scores.
For example, in the equation
Reading Score (ReadScr) = ao + b1 YrsExp + b2 AdvDeg + b3 YrsExp*AdvDeg
+ b4 X(Other X Variables)
the effect of YrsExp on ReadScr is the partial derivative of ReadScr with respect to YrsExp
= b1 + b3 (AdvDeg defined at some level). In our equation, this equals –0.61 +1.5 (AdvDeg). The effect of YrsExp on ReadScr is positive when the school has an average of 40% or more of teachers with advanced degrees (.61/1.5 =.40). This implies that teacher's years of experience do the most to increase test scores in a school environment where there is a high (40% or more) number of teachers with advanced degrees. The positive and significant coefficient on the interaction tells us that these two teacher inputs are complements that work together to improve reading and math scores. These results suggest, also, that more years of teacher experience is associated with lower test scores in schools with less educated faculty (schools with less than 40% of teachers with advanced degrees).

A similar analysis of the effect of AdvDeg on ReadScr shows that test scores are higher in schools with more highly educated faculty, only when there is above average teacher experience (an average of 14.9 years or more 22.4/1.5=14.9). More teacher education appears to be effective at raising scores only when combined with an environment of higher than average teacher experience. In summary, we find that students in schools with less than average levels of teacher experience and advanced education are predicted to have statistically lower scores as compared to students in schools with higher than average levels of these resources. Further, increasing teacher education or experience doesn't help unless there is above average levels of these resources to begin with.

[2] The effect of AdvDeg on Math Score for A schools that have 15 years of teacher experience is –229.09 +16.9 (15) = 24.9 and for non-A schools as -32.18 + 2.16 (15) = 0.24.

[3] The formula for calculating the impact of YrsExp or AdvDeg interaction term is slightly more complicated for a probit function (Norton et al., 2004), however, the interpretation of the effects are quite similar. To calculate the effect of YrsExp on the probability of attaining an A rating (SchA = 1), for example, we find the partial derivative of Prob (SchA=1) with respect to YrsExp.

Recall that the probit function uses the standard normal density function to estimate the impact of the independent variables on the probability of the school attaining an A rating.

The conditional mean of this formulation is

E[SchA|ReadScr, AdvDeg, X] = Φ(ao + b1 YrsExp + b2 AdvDeg + b3 YrsExp*AdvDeg + b4 X(Other X Variables)) = $\Phi(u)$

Where $\Phi(u)$ is the standard normal cumulative distribution function. The effect of YrsExp on Prob (SchA=1) is calculated as follows.

Partial Derivative $\Phi(u)$ / Partial Derivative YrsExp = (b1 + b3 (AdvDeg)) $\Phi'(u)$

Where $\Phi'(u)$ is the standard normal density function, evaluated at the means of the data.

In our equation, this equals(–0.246 + .586 (AdvDeg))* $\Phi'(u)$. Because $\Phi'(u)$ is always a positive number, this still implies the effect of YrsExp on Prob (SchA=1) is positive when the school has an average of 41% or more teachers with advanced degrees (.246/.586 =.41). More teaching experience at a school increases the likelihood a school attains an A rating as long as there is also a high proportion of teachers with advanced degrees (at least 41% of teachers have advanced degrees). We find that non-A schools are those most likely to have a below average number of teachers with advanced degrees and in these schools, years of experience is not necessarily a positive school attribute.

The effect of AdvDeg on Prob (SchA=1) is (-9.52 + .586 (YrsExp))* $\Phi'(u)$. We find that AdvDeg has a positive effect on the likelihood of a school attaining an A Rating only when teachers have an average of 16.2 years of experience (or longer). The results show that more highly educated teachers increase school ratings only if teachers have a significant amount of teaching experience, as well. Schools with below average levels of teacher education and experience are less likely to earn an A and raising either of these measures does not help to increase the likelihood of an A-rating.