A PRELIMINARY ASSESSMENT OF THE POTENTIAL FOR
SYSTEM-FRIENDLY K-12 REFORM

By

John Merrifield*

Abstract

Empirical studies of various K-12 performance measures speak to the potential for improving the performance of the present system. They can address the following question. Given the fixed factors – the basic funding and governance processes that all fifty states share – what are the plausible maximum predicted values of the academic performance measures?

For each of several econometric models, this review computes two predicted academic outcomes with highly favorable explanatory variable values. It addresses the following question: if highly favorable independent variable values (school inputs, socioeconomic factors) had been the average values, how good would the average academic outcome be? It also addresses a second version of that question: leaving the control variables at their mean values, what is the predicted level of academic performance with just highly favorable values for the policy-sensitive variables.

The general finding from this review of econometric studies of K-12 academic outcomes is that the current K-12 system combined with highly favorable policies and socioeconomic conditions would still only produce mediocre results. The conclusion discusses the findings in the context of the K-12 reform debate.

Paper Presented at the 2006
Western Economics Association Annual Conference
San Diego, California

*  Professor of Economics, University of Texas – San Antonio (78207), Senior Research Associate, Fraser Institute, and author of Global Lessons (2005), and The School Choice Wars (2001). Office: 210-458-2519. FAX: 210-458-2515. E-mail: John.Merrifield@utsa.edu; Web: faculty.business.utsa.edu/jmerrifi/

JEL: I21
A PRELIMINARY ASSESSMENT OF THE POTENTIAL FOR SYSTEM-FRIENDLY K-12 REFORM

I. Introduction

The numerous empirical models that connect student test scores to explanatory variables like teacher training and funding levels also speak to the current K-12 system’s ability to achieve significantly improved performance. The aim of this paper is to exploit each fully specified model’s story about improvement prospects through reforms that retain current governance and funding processes; the key fixed factors not represented by explanatory variables.

Since the major policy issue is aggregate achievement level, I excluded studies of performance change and individual student performance. The next section uses highly favorable independent variable values (definition varies with available data) to solve the regression equations for the predicted average level of performance. It also makes that calculation with control variables at their median or mean values, and only highly favorable policy values. Section 3 assesses the predicted maximum values of the performance variables, and discusses interpretation issues. Section 4 discusses the implausibility of the highly favorable policies.

II. Predicted Maximum Student Performance

I discuss the qualifying studies in alphabetical order by the lead author’s last name. Blair and Staley (1995) used data for 266 Ohio urban school districts. After multi-collinearity issues eliminated some plausible variables, the model became:

\[
\text{SCORE} = 20.98 + 0.34\text{SAL} - 0.13\text{PTRATIO} + 0.15\text{AGI} - 3.73\text{MINOR} \\
- 17.28\text{ADC} + 0.43\text{NEIGHBOR} - 0.8\text{COUNT}
\] (1)

Where (avg +/- 2 s.d. = highly favorable value in parenthesis):

\[
\text{SCORE} = \text{a three-year (4th, 6th, and 8th grade), three-test (reading, math, and language arts) composite test score ranging from 0 to 100.}
\]
SAL = Average Teacher Salaries in Thousands (38.53)
PTRATIO = Pupil-Teacher Ratio (14.4)
AGI = Family Adjusted Gross Income, Thousands (57.35)
MINOR = % Minority (0 > avg - 2StdDev)
ADC = % getting Aid to Families with Depdt Children (0 > avg - 2StdDev)
NEIGHBOR = Average SCORE of Contiguous School Districts (66.7)
COUNT = Number of Adjacent School Districts (9.66 = avg + 2StdDev)

To avoid the bias that might result from excluding marginally insignificant variables, each predicted academic performance calculation excludes only variables [COUNT in (1)] with an implausibly signed coefficient. Substituting the highly favorable values into all of the independent variables with plausibly signed coefficients yields a predicted SCORE = 69.5. With the control variables AGI, MINOR, and ADC set at their average values, and only the policy variables set at their most favorable values, predicted SCORE = 58.7.


\[
PSSA = 1,259.86007 + 0.00735OEPPS - 2.37297LI - 0.00012ENR + 0.00086ATS + 1.45744SPCT - 20.87379DLMA + 1.24332JU \quad (2)
\]

Where (Most favorable = actual max or min):

OEPPS = Operating Expenditures per Pupil Spending ($13,170)
LI = Percentage of Low Income Students (3.9)
ENR = District Enrollment (1715)
ATS = Average Teacher Salary ($64,338)
SPCT = Students Per Classroom Teacher (12.0)
DLMA = Whether or not the district is located in a metro area (0)
JU = District gives its teachers the option of joining a union (1)

With only most favorable values, the predicted average PSSA score is 1403.8 out of a total possible of 2200 to 2500;¹ roughly in the middle of the range of scores Pennsylvania now defines as proficient. With only the policy variables set at their most favorable values, the predicted PSSA is 1328.6; in the lower end of the range Pennsylvania currently defines as proficient.

¹ The total possible varies. The cut scores that define the ranges were not available for 1999. 1403 and 1329 are in the current range defined as proficient.
Mensah et al (2005) used 2000-2001 New Jersey school district data to specify models for per pupil expenditure and test scores at various levels. Since the most important test score is the one for the oldest children, I report only the TEST3_AV model; district average scores for grade 11 combined language and mathematics.

\[
\text{TEST3\_AV} = 5.223 + 0.01\text{YR01} - 0.022\text{POOR} + 0.002\text{STU\_POP} + 0.071\text{EXPPP} \\
+ 0.195\text{CS\_INST} - 0.171\text{CS\_ADMIN} - 0.302\text{CS\_OPMAIN} \\
+ 1.124\text{CS\_EXTCUR} - 0.054\text{ABBOT\_DIST} \\
- 0.022\text{LOWINC} - 0.150\text{COSTIDX} + 1.325\text{ATTD\_RATE}
\]  

(3)

Where (in parenthesis, highly favorable values = observed max or min):

- \(\text{YR01}\) = Dummy variable that equals one for year 2001.
- \(\text{POOR}\) = Dummy variable that equals one if the proportion of \(\text{STU\_POP}\) in the school district receiving federal meal aid is over 25 percent (0).
- \(\text{STU\_POP}\) = Number of students enrolled in the school district, expressed in natural log terms (41,378 in 2000).
- \(\text{EXPPP}\) = Natural log of expenditures per-pupil, defined as total operating expenditures divided by average daily enrollment ($13,981 in 2000).
- \(\text{CS\_INST}\) = Cost share for instruction, computed as total instructional spending divided by total operating expenditures (0.73).
- \(\text{CS\_ADMIN}\) = Cost share for administration, computed as total administrative expenditures divided by total operating expenditures (0.06)
- \(\text{CS\_OPMAIN}\) = Cost share for operating and maintenance, equals total operating and maintenance spending divided by total operating expenditures (0.07).
- \(\text{CS\_EXTCUR}\) = Cost share for extracurricular activities, computed as total extracurricular expenditures divided by total operating expenditures (0.06).
- \(\text{ABBOT\_DIST}\) = Dummy variable = 1 for Abbott school districts.
- \(\text{LOWINC}\) = Proxy for the income level of the district, measured by the pct of \(\text{STU\_POP}\) in the school district receiving meal aid under the federal school lunch program (0).
- \(\text{COSTIDX}\) = Geographic cost of living index for the school district, expressed in natural log terms (0.90).
- \(\text{ATTD\_RATE}\) = Average class attendance rate of \(\text{STU\_POP}\) in the school district, in natural log terms (0.97).
With highly favorable values for the independent variables with correctly signed coefficients, the predicted TEST3_AV is 437 out of 450. Independent variable outliers can produce unrealistic results, so where the median plus or minus two standard deviations is less favorable, I also use them to calculate TEST3_AV. That adjustment lowers the predicted TEST3_AV to 403. With just favorable values for policy variables and median values for the control variables POOR, STU_POP, LOWINC, COSTIDX, and ATTD_RATE, the predicted TEST3_AV score for New Jersey is 426 out of 450. With the control variables set at their median values, and only the policy variables set at their highly favorable values (median + or - 2 standard deviations), the predicted TEST3_AV score for New Jersey is 368 out of 450 (81.8%).

Nelson et al (1996) specified equations for statewide average 1995 SAT (Math + Verbal) scores and fourth grade reading scores from the 1994 National Assessment of Educational Progress (NAEPRead4). The highly favorable value is twice the mean or half the mean.

\[
SAT = 1038 - 5.9TT + 0.0424TT^2 + 0.0023PCI + 0.588PRIV - 0.0005PUP - 0.23MINOR - 0.278U + 0.357BARG + 0.528CONFER - 9.61SOUTH \quad (4)
\]

Where (highly favorable value in parenthesis):

- TT = % that take the test (17.6)
- PCI = per capita income ($41,916)
- PRIV = private HS graduate percentage (19)
- PUP = $$ per pupil expenditure ($10,618)
- MINOR = % minority (10.3)
- U = % Urban (34)
- BARG = Collective Bargaining Percentage (100)
- CONFER = Teacher – District Meet and Confer Agreement (100 – BARG)
- SOUTH = 1 for southern region, 0 otherwise (0).

\[
NAEPRead4 = 195.4 - 0.04U - 0.466PRIV - 0.001PUP + 0.026LEP - 0.271ABS + 0.001PCI - 0.136MINOR + 0.0933BARG + 0.125CONFER + 0.202CS<25 \quad (5)
\]

Where (highly favorable value in parenthesis):

- ABS = Absence Rate (0)
- CS<25 = % of Class Sizes under 25 (100)
- LEP = % of Pupils with Limited English Proficiency (0)
- PRIV = private HS graduate percentage (5.1)

---

2 For details, see: http://www.nagb.org/pubs/readingbook.pdf
Substituting the highly favorable values into all of the independent variables with plausibly signed coefficients yields predicted SAT and NAEPRead4 statewide average scores of 1079 and 260 (proficient), respectively. With the control variables set at their mean values, and only the policy variables set at their most favorable values, the predicted SAT and NAEPRead4 statewide average scores are 949 and 232 (basic), respectively.

Sander and Krautman (1991) used 1986-87 American College Test (ACT) scores for Illinois counties to assess the importance of spending and some schooling variables.

\[
\text{ACT Score} = 19.7 - 0.0002SE + 0.00004SS + 0.00002TSA + 0.004TSC - 0.07TE + 0.02PTR - 0.024PT + 0.001URB - 0.00034DEN + 0.000084FI + 0.07CP – 0.15FHF
\]

Where (Most favorable = mean +/- 2 standard deviations):

\[
\begin{align*}
SE &= \text{Schooling expenditures/pupil ($4261)} \\
TSA &= \text{Average annual teacher’s salary ($30,721)} \\
TSC &= \text{Teacher’s schooling - % with master’s degree (59.3%)} \\
CP &= \text{% of the population 25 years or older with a college degree (20.4%)} \\
TE &= \text{Teacher’s experience (17.4)} \\
PTR &= \text{Pupil-teacher ratio (11.4)} \\
SS &= \text{School size (1581)} \\
FI &= \text{Median Family Income ($26,010)} \\
FHF &= \text{Female-headed family (4.0%)} \\
PT &= \text{Percent taking (40.6%)} \\
DEN &= \text{Density (0 > Mean – 2sd)} \\
URB &= \text{Urban (94.2%)}
\end{align*}
\]

After excluding implausibly-signed SE, and solving for ACT with the most favorable values, the predicted average ACT score is 21.8 out of a total possible of 36, or 19.2 out of 36 with just the policy variables set at the most favorable level.

**III. Assessment**

Table 1 summarizes the previous section’s findings. Even if the average student is highly advantaged, and the schools have generous funding and highly qualified teachers, the average level of academic performance is nothing to get excited about. The Mensah results record high predicted outcomes, but some state tests over-state achievement. Noted education historian Diane Ravitch (2006; p 58) notes that, “school officials and editorial writers know by now not to
trust local or state claims of progress until the NAEP results for the states are released every other year, thus verifying or rejecting state claims.” Harvard’s Paul Peterson (2006; p 12) notes that, “the best available yardstick, NAEP is widely considered to be the nation’s report card.” The 2000-2001 New Jersey exams (Mensah, 2005) may be too easy.

The lower predicted state test scores of Blair and Gamrat are more consistent with Nelson’s NAEP results, and the SAT and ACT results of Nelson and Sander, respectively. According to the Blair and Staley model, D/C- average outcomes would result from highly favorable values for all explanatory variables, and ‘flunk’ is the average outcome for average socioeconomic conditions with highly favorable policy values. Nelson’s NAEPRead4 model indicates that average socioeconomic conditions combined with highly favorable policy values only yields an average outcome of “basic” competency (a score of 46.4%).

The NAEP results are seen as more reliable, but the NAEP exams are still criticized for being too easy. For example, Loveless (2004; p 12) notes that:

“The problem solving items on NAEP are not very challenging—at least not in the arithmetic required to answer them. Content taught in first and second grades is at least two years below grade level for fourth graders. But that is the level of difficulty of more than four out of ten (43.6%) problem solving items on NAEP. The second surprising finding is that even though the NAEP items are so easy, fourth graders do not do very well on them.” “The eighth grade items are only slightly more difficult than those on the fourth grade test (3.4 mean grade level). Almost four out of ten items (39.6%) address arithmetic skills taught in first and second grade—six years below the grade level of eighth graders taking the test. Indeed, more than three-fourths of the items (33/43) are at least four years below grade level, taught in the fourth grade or lower. Yet the percentage of eighth graders answering problem solving items correctly is an unimpressive 41.4%. Problem solving items on the eighth grade NAEP only require knowledge of very simple arithmetic. Despite this, eighth graders have trouble getting them right.”

Still, the NAEP scores are not that high, even for the lower grades (e.g. Nelson’s 1994 NAEPRead4) where U.S. students are closer to international norms. Even SAT scores have become suspect for recent changes in what they aim to measure, and for becoming easier in
1994, and harder again in 2005 (Kahn, 2006). Note that the easier post-1993 SAT was the basis of the Nelson SAT model’s disappointing prediction for highly favorable conditions.

An additional NAEP model bolsters the conclusion that with present fixed factors held constant there is a low upper limit on potential academic gains.

Johnson (2000) studied fourth and eighth grade math (1996 data) and reading (1998 data) NAEP data to determine the importance of teacher training. Johnson’s results were not included in section 2 because his observations were individual students rather than district or state average performance. I computed the predicted NAEP values for highly favorable conditions only for the eighth grade results.

\[
\]

\[
\text{NAEPMATH} = 253.2074 - 33.1248BC - 17.7678HP - 2.8480ONWC + 14.3722PC + 4.9266ARMH - 9.6302PPLP - 0.2478GM + 1.6916BDE + 5.9142BDEL + 9.3040ADEL + 4.8690BDOS + 3.9020ADOS \tag{11}
\]

Where:

- BC = Black Communities (0)
- HP = Hispanic Communities (0)
- ONWC = Other Non White Communities (0)
- GM = Gender = Male (0)
- PC = Parents Attended at Least Some College (1)
- ARMH = Students with Additional Reading Material in Home (1)
- PPLP = Students participate in the Free/Reduced Price Lunch Program (0)
- BDE = Teacher Bachelor’s Degree in Education (1)
- BDEL = Teacher Bachelor’s Degree in Education or Language Arts (1)
- ADEL = Teacher Advanced Degree in English or Language Arts (1)
- BDOS = Teacher Bachelor Degree in Other Subjects (1)
- ADOS = Teacher Advanced Degree in Other Subjects (1)

Solving for the NAEP reading score with the most favorable values, the predicted average NAEP score is 281 (56.2%); the minimum score of NAEP’s proficient level for eighth grade reading. Solving for the NAEP math score with the most favorable values, the predicted average NAEP score is 298, just short of the minimum score of NAEP’s proficient level for
eighth grade math. Again, highly favorable conditions yield only mediocre performance.

Because all of Johnson’s explanatory variables were binary, it didn’t make sense to produce the second set of predicted values with only favorable policy-sensitive variables, and with average values for the socioeconomic variables.

IV. Attaining Highly Favorable Values

Increasing average values of the Blair and Staley (1995) explanatory variables to their most favorable values would mean a 25.1% rise in the average teacher salary, and a 20.8% drop in class size. By themselves, each would be quite costly, but combined (more teachers at a much higher salary), probably prohibitively costly. Likewise, the Gamrat model predicted average test score assumed an 81.5% increase in per pupil expenditure, and a 28.1% drop in class size; again, probably prohibitively costly strategies. The Mensah model predicted result of highly favorable conditions assumed a 65.1% rise in average per pupil expenditure, and the Nelson model assumed doubling per pupil expenditures. The Sander and Krautman predicted average test score assumed a 30% increase in average per pupil expenditure.

V. Summary and Concluding Remarks

Highly favorable, probably mostly unattainable independent variable values mostly yield mediocre results; test scores indicating just adequate average academic performance, or less. Even though the most trusted indicator tests, NAEP, are widely seen as easy, highly favorable conditions within the current system would still not cause students to generally ace them. Indeed, except for the New Jersey model, the absolute level of the predicted scores is quite low. In the models and in the real world, attainable values of policy variables produce inadequate performance; ‘Nation at Risk’ performance.
Many people will see these results as a confirmation of the longstanding ‘schools don’t matter’ finding. But the right conclusion to draw from econometric findings in which policy variables have small or insignificant coefficients is that those variables don’t matter much in the context of the fixed background factors that all of the observations have in common. So, the coefficient of a policy variable indicates how much it affects performance now, but not necessarily how much it could affect performance with different underlying fixed factors.

Econometric analysis explains variability in the dependent variable, but not the average level of the dependent variable. Without variability in key fixed factors like governance and funding practices, empirical measurement of their importance to academic performance is not possible. School characteristics don’t matter much now, but they could matter. Basic economic principles say they should matter more than they do now. Repeated urgent calls for reform say that we need to find a way to make school characteristics matter, or create new fixed factors that will substantially elevate the average level of test scores around which non-school factors explain most of the variability. Certainly we need to use research to develop a strong case for what those new fixed factors should be.

References


3 There is a 1966 federal report that many scholars consider the most famous study of education, *Equality of Educational Opportunity*, otherwise known as the Coleman Report for the sociologist James Coleman. It was widely interpreted as finding that schools have no significant effect on student learning.


<table>
<thead>
<tr>
<th>Study (Lead Author)</th>
<th>All Variables</th>
<th>Only Policy Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Most Favorable</td>
<td>Highly Favorable</td>
</tr>
<tr>
<td>Blair</td>
<td>69.5%</td>
<td>N.A.</td>
</tr>
<tr>
<td>Gamrat²</td>
<td>≤63.8%</td>
<td>N.A.</td>
</tr>
<tr>
<td>Mensah</td>
<td>97.1%</td>
<td>89.6%</td>
</tr>
<tr>
<td>Nelson SAT</td>
<td>67.4%</td>
<td>N.A.</td>
</tr>
<tr>
<td>Nelson NAEP</td>
<td>52.0%</td>
<td>N.A.</td>
</tr>
<tr>
<td>Sander</td>
<td>60.8%</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

¹If any. The official possibilities are below basic, basic, proficient, and advanced
²≤“ because the calculation is based on the 2200 low end of the range of the total possible.