Assessing the Costs of K-12 Education in California Public Schools
Technical Appendix

This appendix provides additional detail about several issues raised in the main body of the report: the use of the graduation rate as an alternative outcome measure in the cost function, the influence of Los Angeles Unified in the data, the calculation of pupil weights from the cost function coefficients, and the bias from omitted variables in both the cost and production functions.

Alternative Outcome Measures

The main report focuses on the results from a cost function for California in which the measure of performance is the Academic Performance Index. As shown in Table 3 in the main report, results are quite similar when the API is replaced with percent proficient on the California Standards Test in math and English Language Arts. I also estimated a specification in which performance is measured by the four-year graduation rate. The calculation of this variable is based on the one-year dropout rates for grades nine through twelve and roughly estimates the probability that a student will not drop out during her high school years.[1] Note that this measure of performance only applies to high schools so the sample excludes all elementary districts. This measure is also likely to be a large underestimate of the actual percentage of students who do leave sometime between ninth and twelfth grades (see Losen and Wald, 2005, for a discussion of the problems with using the cumulative one-year measure).

---

[1] Specifically, the four-year graduation rate is equal to one minus $((1-(d9/e9))*(1-(d10/e10))*(1-(d11/e11))*(1-(d12/e12))))*100$, where d# is the number of dropouts in each grade and e# is the number of students enrolled in each grade. Thus, the four-year rate is a cumulative calculation of the one-year rates, rather than a longitudinal measure based on following a cohort over time.

The results from the graduation rate specification is shown in table 3A. For ease of comparison, the first two columns show the results with the average API and CST measures. The coefficients in the specification with the graduation rate are noticeably different in a few ways, though this is also not particularly surprising, given the change in the sample. The coefficients on poverty and regular special education both noticeably decrease while the coefficients on non-Spanish-speaking English Learners and severely disabled both noticeably increase. However, we should keep in mind that there is also likely to be much more noise in the dropout variable; therefore, while it is reassuring that the overall pattern is consistent with the models using test score measures, the results for this specification should be taken as more suggestive than definitive.

Influence of Los Angeles Unified

With almost 700,000 students in 2004-05, Los Angeles Unified is significantly larger than any other district in the state. Given that the cost function is pupil-weighted (giving relatively more weight to larger districts), it is possible that L.A. alone could be skewing the results. To check this, Los Angeles was dropped from the analysis and the regression re-estimated. The results are shown in the fourth column of Table 3A. Although all the coefficients shift slightly, the only change of any noticeable magnitude is the effect of district size. When Los Angeles is included, the coefficients on enrollment and enrollment-squared indicate that there are economies of scale up to a district enrollment of 28,992; above that point, costs begin to rise again. When Los Angeles is excluded, the bottom of the 'U' drops to 19,280.

Calculation of Pupil Weights

To generate a pupil weight, say for poverty, the cost function is used to calculate a base cost per pupil for each district, under the assumption of no students in poverty. That is, all of the other district characteristics are held equal to their original values but the percent of students in poverty is set equal to zero. By multiplying this prediction by the number of students in each district, this gives us a base prediction of total expenditures for each district in the absence of poverty. Next, the cost function is used to calculate per-pupil cost for each district at the actual level of poverty, i.e., the original predicted costs discussed in Table 4, in which percent of students in poverty was set equal to each district's actual value. Again, this is multiplied by the total number of students to get total, rather than per-pupil, spending. The difference between these two predictions is divided by the number of poor students to arrive at a measure of additional cost per student in poverty:

$$\text{Marginal Cost of a Poor Student} = \frac{(\text{Total Spending, Actual \% Poverty}) - (\text{Total Spending, Zero Poverty})}{\text{Number of Students in Poverty}}$$

A district-specific poverty weight is then determined by dividing this marginal cost by the base per-pupil cost for each (non-poor) student. For example, if the marginal cost in a particular district is $1000 and the district's base cost is $5000, then the district's poverty weight is 0.2 (=1000/5000), implying that poor students require twenty percent more resources. This procedure can also be repeated to generate a weight for ELs, special education, or any other variable in the cost function. See Duncombe and Yinger (2005) for a full discussion of using cost functions to generate pupil weights. Although this yields a different weight for each district,

these are typically clustered fairly tightly around the average[2] so it would not be inappropriate to use the average value as a state-level weight.

Omitted Variables Bias in Cost Function

If there are factors that are correlated with spending and student outcomes but that are not explicitly included in the cost function estimation, this can create systematic error in the coefficients. For simplicity, let us assume that spending is a function of two variables, test scores and inefficiency. If we could measure inefficiency, we would estimate the following model:

(i) $S = \beta_0 + \beta_1 T + \beta_2 I + u_s$

where β_1 captures the 'true' relationship between spending and test scores, β_2 captures the correlation between spending and inefficiency, and u_s is random error. Since we cannot observe inefficiency, we instead estimate a model with just test scores:

(ii) $S = = \beta_0' + \beta_1' T + v$

where $v = \beta_2 I + u_s$. If we estimate this model with only test scores, we get the estimated coefficient, $\hat{\beta}$.

Now assume that test scores and inefficiency are also correlated and we can express that by setting up inefficiency as a function of test scores:

(iii) $I = \delta_0 + \delta_1 T + u_i$

so δ_1 represents the relationship between inefficiency in test scores. If we plug (iii) into (i), we get:

(iv) $S = \beta_0 + \beta_1 T + \beta_2(\delta_0 + \delta_1 T + u_i) + u_s = \beta_0' + (\beta_1 + \beta_2 \delta_1)T + u$

---

[2] For example, the district-specific poverty weights in California range from 0.28 to 0.32, around the mean of 0.3.

Thus, if spending is regressed solely on test scores, as in (ii), the expected value of $\hat{\beta}$ is not equal to the true value, β_1, but instead is equal to $(\beta_1 + \beta_2\delta_1)$. β_2 is the correlation between spending and inefficiency and δ_1 is the correlation between test scores and inefficiency. If either of these is zero, then the bias goes away and the expected value of $\hat{\beta}$ is the true value β_1. If one correlation is positive and the other is negative, then $\hat{\beta} < \beta_1$. If both correlations are positive, then $\hat{\beta} > \beta_1$.[3] The size of the bias will depend on the strength of these two correlations.

Production Function as the inverse of the Cost Function

The cost function is estimated with spending as a log-linear function of test scores, teacher salaries, and student and district characteristics:

$$ln\ S = \beta_0 + \beta_1 ln\ API + \beta_2 Sal + \beta_3 Student\ Char + \beta_4 District\ Char + u_s$$

In a production function, outcomes are modeled as a function of spending and the other variables:

$$ln\ API = \alpha_0 + \alpha_1 ln\ S + \alpha_2 Sal + \alpha_3 Student\ Char + \alpha_4 District\ Char + u_T$$

In theory, the cost function coefficient on student performance, and the production function coefficient on spending would be inverses of each other. That is, using the above notation, $\alpha_1 = 1/\beta_1$. If the data and model were perfect (i.e., correctly specified with no unobservable variables), estimates from the cost function and production function should be similar. However, if spending and test scores are both correlated with unobservable inefficiency, then the estimates in both cases will be biased, following the logic outlined in the previous section.

---

[3] Note that the analysis is significantly more complicated when there are more independent variables, as there are with the California cost function. However, the intuition about the direction of the bias is generally the same (Wooldridge, 2006).

Note that if the omitted variable is positively correlated with spending and negatively correlated with test scores, the bias will lead the estimated coefficients to be smaller than their true values. Given that the relationship between spending and test scores appears to be positive ($\hat{\beta} > 0$, $\hat{\alpha} > 0$), this means that the estimated coefficients will be closer to zero. This creates a serious problem because a small coefficient on test scores in the cost function and a small coefficient on spending in the production function have opposite implications. A small coefficient on test scores in the cost function indicates that a large change in test scores is associated with a small change in dollars; this could be interpreted as implying that dollars are extremely effective. In contrast, a small coefficient on spending in the production function indicates that a small change in dollars buys only a small change in test scores; this could be interpreted as implying that dollars are *not* very effective. For a district to achieve any given test score level, the cost function would predict a much lower spending level than would the production function. Without the bias, the two estimates should be quite similar. Thus, the fact that the cost function estimates are so much smaller than the production function estimates is one indication that the bias does, indeed, exist, and that the magnitude is not negligible.

Additional references:

Losen, D. and Wald, J. (2005). Confronting the Graduation Rate Crisis in California, Civil
       Rights Project, retrieved August 21, 2006 from
       http://www.civilrightsproject.harvard.edu/research/dropouts/dropouts05.pdf.

Wooldridge, J.M. (2006). *Introductory Economtrics: A Modern Approach* (Mason, OH:
       Thomson Higher Education).

Education Cost Function, 2004-05
California K-12 School Districts

Dependent variable: Log of expenditures per pupil

Independent variables	API	CST	Graduation Rate	API - without LA Unified
2004-05 Performance	0.339*	0.308*		0.324*
	[2.29]	[2.63]		[2.21]
Four-year Graduation Rate			0.001	
			[0.72]	
Teacher Salary Index	0.321**	0.296*	0.331**	0.346**
	[2.95]	[2.61]	[3.33]	[3.29]
Log of student enrollment	-0.286**	-0.266**	-0.351**	-0.371**
	[11.24]	[10.84]	[10.10]	[6.19]
Square of log of student enrollment	0.014**	0.013**	0.017**	0.019**
	[12.31]	[11.88]	[10.44]	[5.80]
Percent of students eligible for free and reduced price lunch	0.283**	0.344**	0.219**	0.265**
	[3.78]	[3.60]	[4.43]	[3.31]
Percent of students Spanish-speaking English Learners	0.081	0.078	0.013	0.098
	[1.09]	[1.06]	[0.16]	[1.22]
Percent of students non-Spanish-speaking English Learners	0.237	0.204	0.468*	0.203
	[1.27]	[1.08]	[2.38]	[1.12]
Percent of students in special education	1.070**	1.351**	0.664*	1.170**
	[4.40]	[5.30]	[2.34]	[4.98]
Percent of students with high-cost disabilities	6.506**	5.685*	8.897**	6.143**
	[2.98]	[2.47]	[4.16]	[3.04]
Percent of students enrolled in high school	0.247**	0.232**	0.211**	0.240**
	[7.89]	[5.03]	[6.89]	[7.15]
Herfindahl (efficiency) index	-0.02	-0.06	0.04	-0.06
	[0.20]	[0.45]	[0.35]	[0.51]
Intercept	7.423**	9.434**	10.019**	7.890**
	[7.07]	[41.66]	[46.43]	[7.44]
Observations	937	808	407	937

** indicates statistically significant at the 5% level