Alternative Assessments of the Performance of Schools

Measurement of State Variations in Achievement

Eric A. Hanushek
Lori L. Taylor

ABSTRACT

Evaluation of the efficacy of school policies requires measures of student performance across schools and states, but conventional approaches to constructing the relevant data can be very misleading. This paper develops an approach to estimating marginal school effects at the state level. It then documents and estimates the magnitude of biases introduced by commonly employed estimators of school quality. Direct estimates of achievement growth, or value-added, are shown to be far superior to any alternative correction that is commonly employed. Especially at the state level, nonrepresentative data such as aggregate SAT scores provide very biased measures of school quality differences—even when statistical adjustments for demographic differences and varying participation rates are employed.

I. Introduction

School reform reports and initiatives of recent years have focused significant attention on state educational policies. The character of teacher certification rules and tenure laws, the potential use of minimum teacher salary scales, and the imposition of uniform graduation

Eric A. Hanushek is a professor of economics at the University of Rochester. Lori L. Taylor is a research economist at the Federal Reserve Bank of Dallas. *This is a substantially modified version of a paper originally presented at the 1987 APPAM Research Conference in Washington, DC. The authors are grateful for helpful comments by Marcus Berliant, Charles Phelps, and anonymous referees.*
[Submitted April 1988; accepted March 1989]

requirements based upon minimum competency examinations typify state level policy issues being currently addressed. But the quality of the evidence that leads to recommendations and conclusions about such state policies has had a distinctly anecdotal character, in large part because data about performance of the schools in different states are not directly available but must be inferred.

Most insights into the effectiveness of alternative educational policies have come from microanalyses relying on student performance in individual schools or classrooms. These studies are, however, inherently limited in their ability to address issues of state policies, since without consistent observations of schools across different states it is impossible to observe the ramifications of statewide policies.

This paper systematically evaluates alternative strategies for estimating variations in the marginal effectiveness of schools across states. While the ultimate objective is an assessment of the impacts of alternative state policies, this analysis stops at the logically prior step of measuring performance differences across states. The usefulness of school evaluations based on data from the Scholastic Aptitude Test (SAT) is also assessed.

II. Available Evidence and Inference

Informed policy making requires being able to relate resources or policies to their marginal impacts on performance. But, no single source regularly provides appropriate data on the performance of schools that can be used for analyzing policies. Virtually the only relevant evidence on performance as related to state differences has come from average scores of college-bound students taking the SATs or the ACTs (American College Testing program), but these data are prone to serious problems in the evaluative context considered here.

Because many factors affect school performance, the ranking of state averages on tests does not provide an appropriate ranking of the marginal differences in school effects by state. Moreover, even if *only* schools were important, the aggregate data might still be quite misleading because of measurement problems and selection bias issues.

Consider a simple model of individual achievement where the current and past patterns of two factors—schools (S) and families (F)—contribute systematically to learning.[1] After acknowledging that some attributes

1. S and F are best thought of as vectors of separate factors influencing eductional performance. S includes, among other things, measures of state policies, the focus of this work. For expositional convenience, this is written as a linear relationship, but this is not fundamental to any of the results.

of individual performance defy measurement, achievement can be written as:

$$(1)\quad A_{iT} = \alpha_T S_{iT} + \beta_T F_{iT} + \sum_{t=1}^{T-1} \alpha_t S_{it} + \sum_{t=1}^{T-1} \beta_t F_{it} + \sum_{t=1}^{T} \epsilon_{it}.$$

A_{iT} is the achievement of student i in school year T. The parameters α_t and β_t are weights attached to school resources and family resources in the various past school years ($t = 1,2, \ldots, T-1$) and in the current school year (T) and the ϵ_{it}'s represent the unmeasured factors that contribute to achievement. The α_t's are simply the marginal effects on student achievement of school inputs in different years, just what must be known for policy judgments.

It is also useful to think of the unmeasured factors as involving two components: (1) a systematic individual specific part (δ_i) that is the same over time for an individual but that varies across students; and (2) a random part that varies over individuals and time (θ_{it}); that is,

$$(2)\quad \epsilon_{it} = \delta_i + \theta_{it}.$$

The systematic individual component represents differences in intelligence, motivation, unmeasured family inputs, and other things that directly contribute to performance—factors often labeled simply individual "ability."

The formulation of Equations (1) and (2) provides a straightforward way of assessing the most common problems arising in the analysis of statewide school performance, an exercise that invariably involves only publicly available aggregate data. The prototypical analysis begins with a sample of students in each state (s) who have taken an achievement test. The average state test performance (A^s) is then related to readily available state averages for current levels of school resources (S^s)—measured by, for example, average school expenditures—and family inputs (F^s)—measured by, for example, average state income. This situation can be depicted by a model such as Equation (3):

$$(3)\quad A^s = a_T S^s + b_T F^s + e^s.$$

The key to what might go wrong, of course, is that Equation (3) differs from the way the data were generated [Equations (1) and (2)]. In particular,

$$(4)\quad e^s = f^s(S_1, \ldots, S_{T-1}, F_1, \ldots, F_{T-1}, \theta_1, \ldots, \theta_T, \delta)$$

where the arguments of the function are state averages of the previous school and family factors plus the aggregate individual error terms—the

things left out of the estimation.[2] The estimates from Equation (3) depend upon the correlation of e^s with S^s and F^s, on the structure of the average error terms in Equation (4), and on the temporal structure of school and family inputs.

The serendipitous case would be a situation where the contemporaneous school measure (S^s) is perfectly correlated with all of the relevant past school data for the individuals and both S^s and F^s are uncorrelated with the other elements of the composite error, e^s. The contemporaneous achievement model would then yield reasonable estimates of the marginal effect of school resources across all of the years. Clearly, however, the conditions for such direct interpretability are very unlikely to be met when analysis relies upon sampling and data collection that are outside of the analyst's control.

Estimation based upon available aggregate state level data immediately suggests several likely classes of problems, sketched here. The empirical work then isolates and assesses the quantitative importance of each.

Case 1: Missing Family Data

The most obvious analytical problem arises when a "bivariate" analysis of school differences which ignores family influences is attempted. This polar case corresponds to such casual analyses as correlating state performance differences with differences in expenditures or pupil-teacher ratios, an approach clearly more prevalent in popular and policy discussions than in research work. The error term in the estimation equation implicitly includes both e^s in Equation (5) plus the contemporaneous family factors, F^s. Family educational inputs are expected to be positively correlated with the quality of schools; more educated and higher income parents tend to provide better education at home and to search systematically for better schools. The result is simple: other things equal, the estimated importance of school factors in determining achievement will be overestimated.

Case 2: Time Varying School Inputs

When analysis of educational performance relies upon different data sources, it is practically impossible to measure accurately the historical

2. Of course, aggregation to the state level may induce bias in and of itself. This bias may be particularly problematic for analyses of school effects given the probable correlation between school and family characteristics at the local level. The impact of aggregation bias is estimated separately in the empirical analysis below.

pattern of school inputs that are relevant for the sampled students. Contemporaneous school measures will not reflect historical inputs when there are unsystematic changes in school policy over time (such as those arising from the imposition of state expenditure limits) or when the students themselves migrate from different states. For example, the use of a final school year instrument like the SAT can very easily attribute to the state of twelfth-grade residence school effects which were developed years before and hundreds of miles away. The direction of bias in the estimates of schooling parameters cannot be determined a priori, because, unlike standard measurement error problems, the errors in variables frequently are correlated with the true input values.[3]

Case 3: Measurement Error in School Inputs

Typically, mismeasurement of the inputs will bias the estimated parameters toward zero, or toward finding no relationship. This is important when considering estimated school effects, since the available research on school relationships (see Hanushek 1986) suggests true uncertainty about how to measure school inputs. Additional error comes with aggregate data collected from the mismatched sampling of students and schools, such as SAT or ACT scores by state which make no distinction between students in public and private schools. Public school data simply do not reflect accurately the relevant school inputs for SAT and ACT takers, and the severity of measurement errors might well interact with the nonrandom test taking (see below).[4]

Case 4: Nonrandom Test Taking

Interacting with each of the previous issues is the possibility of selective test taking by students. SAT or ACT scores provide the most obvious examples, since scores are obtained only for students with some thought of continuing on to college. The problem can also occur, however, with

3. The same type of measurement problem arises with family inputs but is likely to be less severe because of the greater stability in family inputs, particularly at the aggregate level.
4. In most discussions, errors of measurement in the exogenous variables are the focus of attention, since the error terms in the estimation equation (ϵ_{it}) can be thought of as including measurement errors in the endogenous variable. With educational performance, however, mismeasurement of the dependent variable can arise from the conditions of test-taking, the nature of the tests themselves, the opportunity for various school systems to "teach to the test," or simply the measurement of the wrong things. Such errors in measuring achievement will increase the uncertainty of any parameter estimates and can lead to systematic bias in the estimated parameters, depending on whether or not the errors are correlated with the inputs to achievement.

other performance measures resulting from, say, incomplete test administration across school districts.[5]

The influence of test taking proportions has been cited as an explanation of changes in aggregate test scores (e.g., Wirtz et al. 1977, CBO 1987). More relevantly for this analysis, because of the geographic pattern of the use of SATs, the proportions taking the test varies dramatically across states: from 2 percent in South Dakota to 69 percent in Connecticut in 1982.

If smarter students in each state tend to take the test, varying proportions of test takers will imply the mean individual component (δ^s) will vary inversely with the proportion taking the test in each state. Since characteristics of families (such as income or education levels) and characteristics of schools also enter into college attendance and test taking, a correlation between δ^s and both S^s and F^s will be induced. This in turn implies biased estimates of the schooling parameters even when school resources and family inputs are well measured. The direction of bias in state level analyses cannot, however, be determined a priori.

The exact nature and magnitude of each of these problems depends, of course, on the specifics of the data and the underlying structure of educational performance. Estimation biases depend upon such things as the aggregate correlations among educational inputs over time or the precise pattern of measurement errors, and available analyses simply tell us little about these things.

This paper provides a systematic investigation of the magnitude of each of these problems within a consistent analytical framework and using a data set which allows empirical investigation of the complicated interactions involved.

III. Empirical Analysis

The empirical analysis employs data provided for the sophomore (10th grade) cohort in the High School and Beyond (HSB) sample. These data allow designing samples and estimation models that isolate the various potential effects previously identified. Following this, comparisons are provided with estimates derived directly from state level SAT and ACT test data.

5. Selection bias of a different sort may arise at the local level due to the ability of parents to choose the schools their children attend through their choice of residence. This effect is of limited significance at the state level, however. With the exception of communities on the border between states, the likelihood of parents selecting their state of residence according to the characteristics of the schools seems particularly low.

To separate the estimation problems discussed above from those created by aggregation itself, the empirical work will rely primarily on individual data. Synthetic aggregates will then be introduced for comparison and for estimation of any aggregation bias.

A. Data Overview

The HSB data set was gathered between 1980 and 1984 by the National Opinion Research Center for the U.S. Department of Education. The portion of the data pertinent to this investigation follows the secondary education of up to 36 students in the 1980 sophomore class from each of 767 public high schools in the United States.[6] This paper makes use of the mathematics, science and vocabulary skills tests administered to the students in the early spring of their sophomore year, and again in the early spring of their senior year. Demographic surveys completed by the students at the same time as the academic tests provide a wide range of information including the student's sex, race, family income, and history of participation in ACT and SAT testing. For a limited subset of students, there are also reports of each student's actual score on SAT and/or ACT tests taken prior to February of the student's senior year. While no information is provided on the specific geographic location of the schools, the state location can be inferred from information in the total data set concerning post-secondary schooling. The inferred location, discussed in Appendix A, is used in this analysis. Only observations with complete demographic data including scores on at least one HSB test and for which a state of residence could be inferred are included in the working data set.[7]

Each of the separate HSB academic achievement tests, while brief, is reasonably highly correlated with the more extensive SAT and ACT instruments. As seen in Table 1, the individual level correlations for the combined math and vocabulary scores is about .85 for both the SAT and

6. The selection of schools included in the sample is not completely random. Private schools and public schools identified as Hispanic are overrepresented. Selection of students within schools was, however, random.

Schools are lost within this analysis because: 1) the state cannot be inferred (see Appendix); 2) they are private; 3) the school did not administer the HSB test battery; or, 4) all students in the sampled school were missing one or more of the key data elements. These reasons combine to reduce the sampled schools from 1,000 to 767. An unknown number of students were lost between the sophomore and senior year because of migration, earlier graduation, illness, failure to complete the tests, and so forth.

7. Given the nature of the procedure used to infer state locations, schools on the border between states are likely to have been excluded from the working data set. This has the effect of virtually eliminating the possibility of selection bias caused by parents choosing their residence according to the characteristics of the schools.

Table 1
Correlations of HSB Tests and SAT/ACT Tests (Numbers of observations in parentheses)

	High School and Beyond		
High School and Beyond	Vocabulary	Math	Science
Vocabulary	1.00	.69	.70
		(25,598)	(25,280)
Math		1.00	.70
			(25,302)
Science			1.00

	High School and Beyond			
HSB Reported SAT/ACT[a]	Vocabulary	Math	Science	Math + Voc
SAT	.73	.78	.68	.85
	(3,206)	(3,180)	(3,137)	(3,175)
ACT	.72	.78	.71	.84
	(2,352)	(2,325)	(2,314)	(2,321)

Note: a. Individual student data for the subset of schools where reports of SAT or ACT test scores were provided within the HSB data set.

ACT scores. The separate tests, while correlated with each other, can also be seen to display considerable independent variation.

B. Empirical Specifications

The analysis of biases in the schooling factors is based on, first, estimating alternative models that deviate in known ways from the fully specified model and, second, creating samples with synthetic sample selection. The base case for all comparisons is a value added model with a school level covariance structure. This can be written as:

$$(5)\quad A_{i,12} = \lambda A_{i,10} + \beta F_i + \sum_{k=1}^{n} q_k SCH_{ik} + \epsilon_i$$

where $A_{i,12}$ is the twelfth grade achievement of the ith student; $A_{i,10}$ is the tenth grade achievement of the ith student; F_i is a vector of family background characteristics pertaining to the ith student; the SCH_{ik} are dummy variables that equal one if the ith student goes to school k and equal to

zero otherwise; ϵ_i is a random error assumed to be orthogonal to the regressors; n is the number of schools in which students are sampled; and λ, the β's, and the q_k's are unknown parameters to be estimated.

Equation (5) is the empirical specification of Equation (1) that is appropriate for estimation using the HSB tests as the performance measure. This specification assumes that all of past educational inputs are captured by the tenth grade score so that estimation of the value added form (Equation 5) eliminates unmeasured school and family factors from the past and will minimize any individual specific differences.[8] Further, the covariance structure avoids problems of measuring the specific school factors that count in the determination of performance. Thus, the specification of Equation (5) avoids the most important measurement and omitted variables problems.[9]

Our purpose here, however, is the development of measures of state educational quality and the assessment of alternative estimation methods. There are two obvious approaches to this. The first simply decomposes the estimate of school quality, q_k into a statewide average component, s_j for state j, and an orthogonal school-specific component (found by taking variations from the state mean for q_k). The set of estimated state schooling parameters $\{s_j\}$ provides a baseline estimate of the variations in school quality across states used in all subsequent comparisons; this set is denoted S-BASE.

This estimator, while close to the theoretical ideal, is frequently impractical in that it requires an underlying school based sample. An alternative is estimation of a simplified state covariance model such as:

$$(6) \quad A_{i,12} = \lambda A_{i,10} + \beta F_i + \sum_{j=1}^{m} s_j' S_{ij} + \epsilon_i'.$$

In Equation 6, only state dummy variables—S_{ij} for the m states—are included, implying that school specific quality differences are included in

8. Boardman and Murnane (1979), beginning with a model similar to equations 1 and 2, discuss alternative interpretations of such value-added estimation. As they point out, if δ affects both the level and the growth of achievement, any components of δ that are orthogonal to prior achievement and the included regressors could bias the parameter estimates. The covariance structure of school effects and the concentration on twelfth grade performance, however, imply any such bias should be small.

9. If the set of relevant school factors were known and measured, the estimation design employed here would be less efficient than simply including the specific factors. It would also provide less information because the alternative would give estimates of the marginal effects of specific factors. On the other hand, the alternative will yield biased and inconsistent estimates if there are remaining measurement and specification problems. This formulation ignores any within school variations in school quality, but for the purposes of estimating state effects this will almost certainly have minimal effects on the estimates.

ϵ'. These estimates, identified as *S-STATE*, might be expected to be close to *S-BASE* since the omitted quality component is orthogonal to the included state effect; if so, this is a more practical estimation scheme that also carries over naturally to aggregate estimates.[10]

The alternative specifications, corresponding to the previously identified problems, are now straightforward. The "bivariate" specification, corresponding to Case 1, is simply:

$$(7) \quad A_{i,12} = \sum_{j=1}^{m} s_j^* S_{ij} + \epsilon_i^*.$$

The new set of state school quality parameters $\{s_j^*\}$, which is equivalent to calculating state means for raw test scores, will no longer reflect just marginal schooling effects but will also incorporate some of the family effects, of the systematic measurement problems, and of the historical measurement errors in school factors.[11] These estimates, which represent the extreme in misspecification, are subsequently referred to as *S-RAW*.

By adding contemporaneous family measures, the specification fits Case 2, and the school estimates (referred to as *S-DEMOG*) are found from:

$$(8) \quad A_{i,12} = \beta^{**} F_i + \sum_{j=1}^{m} s_j^{**} S_{ij} + \epsilon_i^{**}.$$

In the empirical work, the student's sex, race, and family income are used to measure F_i. While a wider range of background measures are available in the HSB data, these correspond to demographic factors typically available and used in adjustments at the aggregate level.

There are many alternative explicit measures of school quality that can be employed, so that the measurement problems of Case 3 come in different forms. Here we concentrate on expenditures per student (EXP_j) as the school quality measure, yielding in a value-added form:

$$(9) \quad A_{i,12} = \lambda A_{i,10} + \beta F_i + \gamma EXP_j + \epsilon_i''.$$

By comparing $s_j^e = \gamma EXP_j$ (which is denoted *S-EXPEND*) to s_j in each

10. Note that with aggregate data the school specific error terms in each state would sum to zero, implying consistent estimates of the parameters in Equation (6). Offsetting this would be any variation in individual or school level parameters within the states that could imply a classic aggregation bias problem; see, for example, Theil (1971).

11. An even more restrictive version would specify precise measures of school differences such as expenditures per student. This then incorporates the mismeasurement of contemporaneous school differences. Some evidence on this version is presented below.

state, the most common version of explicit contemporaneous measures—and the associated measurement errors—can be examined.

Finally, a simplified version of the value added is estimated:

$$(10) \quad A_{i,12} = \lambda A_{i10} + \sum_{j=1}^{m} s_j^{***} S_{ij} + \epsilon_i^{***}.$$

This estimation, which ignores any specific family inputs, is viewed as a shortcut and would be expected to yield results quite close to the full state estimates as long as family background factors are stable over time. These estimates are denoted as *S-PRETEST* in the subsequent analysis.

Three alternative summary statistics are computed—each giving somewhat different information about the relationships among the different quality estimates. First, the Pearson correlation of *S-BASE* with each estimate gives an indication of whether the alternative methodologies produce similar or dissimilar views about the relative effectiveness of schools in the various states.[12] The correlation coefficients do not, however, indicate whether school effects are overestimated or underestimated. This is found by creating a measure of school quality directly from the estimates of *S-BASE*. Specifically, let the value of the constructed school quality variable, *SQ*, for state j simply equal s_j, the mean school effect in state j. If we reestimate Equation (5) [or (6)] in the form:

$$(5') \quad A_{i,12} = \lambda A_{i,10} + \beta F_i + \phi SQ + \epsilon_i,$$

ϕ will equal 1 for an unbiased estimate of the effect of state school quality. When the state dummy variables in the subsequent misspecified models [Equations (7)–(10)] are replaced with *SQ*, deviations of ϕ from one will indicate the direction and magnitude of bias in the marginal effects of schools. The statistic $(\phi - 1)$ will be positive (negative) when there is upward (downward) bias and is referred to as "marginal bias" in the subsequent tables.[13] The estimated standard errors for ϕ can be used to develop approximate tests for bias.[14]

To capture how close or far away the different estimates of state quality

12. In addition, Spearman rank order correlations were calculated. In these only ordinal information about state rankings on quality is used. However, these yield no additional information with samples of this size (45 states), so they are not reported here.

13. The analysis of marginal bias in the case of expenditure measurement is different from the others. The state school effect is $s_j^e = \gamma EXP$ from Equation (9). Estimates of ϕ are then obtained from the auxiliary regression of $S_j^e = a_o + \phi SQ + u$. Marginal bias is always estimated from samples including all states.

14. These tests are correct for large samples, but are only approximate in small samples because they do not take into account the estimation errors in the underlying coefficients *S-BASE*.

are from each other, an index of squared bias (*SB*) is created by comparing the sum of squared deviations of each estimate from *S-BASE* to the underlying variation in *S-BASE;* e.g., for *S-RAW,* this would amount to[15]

$$SB = \Sigma[(s_j - \bar{s}) - (s_j^* - \bar{s}^*)]^2/\Sigma(s_j - \bar{s})^2.$$

This latter index is designed to be an extension of the natural computation of relative squared bias for an individual coefficient to the case where a series of separately estimated parameters reflect the effect of schooling on achievement. When *SB* equals one, the estimation error is as large as the underlying variation in school quality. Larger values of *SB* imply that the errors in estimation of the school quality dominate the observed variations.

C. Basic Results

The concentration on state variations in school quality is unusual, particularly when data on individual schools exist. Yet, state educational policies and state expenditure policies differ dramatically so that concomitant variations in performance would be expected. Of the total variation in school performance (the q_k's), 10 percent is found between states—indicating that aggregate policies do in fact have an impact.[16]

While the preferred estimation method allows for variations in school quality within states, school quality can also be estimated based on state level estimates [Equation (6)] using either individual student (*S-STATE*) or aggregate state (*A-STATE*) data. These alternatives, which each rely on variants of the well specified model of educational performance, are compared for the different HSB tests in Table 2. When estimated from individual student data, there are only minuscule differences in state quality estimates according to the two approaches. The correlation between the two (*S-BASE* and *S-STATE*) is .97 for each test, and the marginal bias is very small. Thus, even though a majority of the variation in school quality is found within states, ignoring such variation has virtually no effect on estimates of state school quality because within state variations are not highly correlated with the measured inputs.

When aggregate data are used (*A-STATE*), the bias (particularly for vocabulary and science achievement) remains small, although the correla-

15. The mean of the school effects estimated in this manner contains the "intercept" of the model and is affected by the other variables in the model. Because of the arbitrariness of the mean value, it is removed from the calculations.

16. When the variation in school quality estimates is decomposed into within state and between state components, the between state part ranges from 9.9 percent of the total variation in math to 12.2 percent in science.

Table 2
School Level Estimates (S-BASE) *versus State Level Estimates* (S-STATE) *and Aggregate Estimates* (A-STATE) *of State School Quality*

Alternative estimates	Vocabulary	Math	Science	Math + Voc
S-STATE (Equation 6)				
Squared bias (SB)	.06	.06	.07	.07
Pearson correlation	.97	.97	.97	.97
Marginal bias ($\phi - 1$)	−.04	−.08	−.03	−.07
(standard error)	(.02)	(.03)	(.02)	(.03)
A-STATE[a]				
Squared bias (SB)	.69	.28	.77	.47
Pearson correlation	.59	.85	.54	.74
Marginal bias ($\phi - 1$)	−.02	−.11	.03	−.11
(standard error)	(.04)	(.04)	(.06)	(.04)

a. Estimates of state school quality using aggregate data according to Equation (6).

tion with the base estimates falls from that for *S-STATE*. The increased imprecision of the school quality estimates undoubtedly reflects the larger sampling errors arising from the small samples for the aggregates (46 states).

Table 3 presents the overall results of the effects of the different common problems (and related empirical specifications). Not surprisingly, the top panel demonstrates that state by state differences in raw scores (*S-RAW*) are very far from the baseline estimates of the marginal effects of schools. The simple correlation with base school quality ranges between .5 and .7, depending upon the specific test.[17] This picture of distortion in the raw test scores is reinforced by the squared bias calculations: the error variance from misspecification is 7–12 times as large as the underlying variation in school quality. The marginal bias ($\phi - 1$), estimated according to misspecifications of Equation (5)′, shows that as expected the naive model based on just raw scores overestimates the marginal effects of schools by a factor of three to five.

17. Four states (Alaska, Delaware, New Hampshire, and Vermont) were eliminated from the analysis of Pearson correlations and squared bias because they had fewer than 25 valid observations. The remaining state data sets range from 31 observations in North Dakota to 1,662 observations in California. The relatively low correlations reported also hold when the estimates are weighted by the number of observations in each state; the Pearson coefficients increase by .03 to .05 after weighting.

Table 3
Effects of Alternative Specifications on Estimates of State School Quality: Comparisons to Fully Specified Estimates (S-BASE)[a]

Alternative Specification	Vocabulary	Math	Science	Math + Voc
S-RAW (Equation 7)				
Squared bias (SB)	7.38	7.62	6.56	12.78
Pearson correlation	.77	.50	.74	.53
Marginal bias ($\phi - 1$)	1.86	3.15	1.47	3.71
(standard error)	(.01)	(.01)	(.01)	(.01)
S-DEMOG (Equation 8)				
Squared bias (SB)	1.97	2.80	1.65	4.14
Pearson correlation	.85	.66	.82	.69
Marginal bias ($\phi - 1$)	1.80	2.96	1.60	3.55
(standard error)	(.02)	(.03)	(.01)	(.03)
S-EXPEND (Equation 9)				
Squared bias (SB)	.84	.91	.69	.90
Pearson correlation	.41	.30	.58	.32
Marginal bias ($\phi - 1$)[b]	−.87	−.79	−.85	−.84
(standard error)	(.004)	(.01)	(.01)	(.01)
S-PRETEST (Equation 10)				
Squared bias (SB)	.18	.14	.19	.14
Pearson correlation	.92	.93	.91	.93
Marginal bias ($\phi - 1$)	−.14	−.22	−.23	−.21
(standard error)	(.01)	(.02)	(.01)	(.02)

Notes: a. For the estimates of Pearson correlations and Squared Bias, only those 46 states represented by more than 25 observations are included.
b. See footnote 10 for description of estimation.

Since the other estimation specifications use additional information in estimating the state effects, it is natural to find that they in fact are closer to the base values. Demographic adjustments for income, race, and sex differences increase the correlations (now .7 to .85) and reduce the squared bias to one quarter of the previous values. They also narrow the distribution of marginal bias, reducing the overestimation of school effects for each test. These effects are precisely what were hypothesized. It is important, however, to note that these corrections come nowhere near eliminating the bias: The marginal effects of schools are still estimated to be at least two and a half times as large as they really are (i.e., the marginal bias ranges from 1.6 to over 3 for the combined math and vocabulary). Ignoring both the history of inputs and variations in ability (δ) leads to significant overestimates in school quality variations.

The third panel, which relies on school quality estimates derived from measuring quality by expenditures per student across states, is perhaps the most interesting. These estimates of school quality (*S-EXPEND*) are biased downwards quite dramatically with marginal bias estimates approximately $-.8$. This is a "pure" school measurement effect since the underlying estimation [Equation (9)] is based on a value-added specification. These estimates are, however, better in terms of both squared and marginal bias than just the demographically adjusted estimates which measure schools better but fail to include historical inputs or individual abilities.

The final panel relates to estimates derived from employing corrections for just tenth-grade performance (i.e., including just the pretest information in addition to the state dummy variables). The estimates of *S-PRETEST* are quite close to those from the base model. Leaving out socioeconomic characteristics of students does lead to some increase in the marginal bias of the school quality estimates (on the order of $-.1$ over *S-STATE*), but this very simple procedure is far superior to either the adjustments that lack pretest information or to the expenditure estimates.

These results clearly indicate the magnitude of bias caused by the different problems. Adding demographic information reduces squared bias to one quarter of that found with just raw score differences, but even larger improvements come from looking simply at gains in achievement. The bias from estimation in level form, which ignores individual ability differences and historical variations in inputs, far exceeds any bias introduced by imperfect measurement of school factors. For example, the squared bias from only a pretest adjustment is just one tenth of that found with an adjustment for demographics that ignores prior test performance. Further, while there are many ways to measure explicitly the quality of schools, using perhaps the most common—expenditures per pupil—shows the dramatic downward bias imparted by mismeasurement.

All results have been derived from analysis of individual data, but using state aggregates yields the same qualitative conclusions. This aggregate confirmation of the findings is important since state level data are more plentiful than national samples of individuals.

D. Nonrandomness

The previous analysis considered alternative ways of estimating school effects using a representative national sample. In contrast, the only data source historically available about schooling across states and over time has been the college testing information of the SAT and ACT programs, where scores are available only for individuals who are motivated to take the test to gain college admission. Since this is a distinctive group whose

Table 4
Effects of Test Selection on Estimates of State School Quality: Comparisons to Estimates with Random Sampling (S-BASE)

Estimation Sample	Vocabulary	Math	Science	Math + Voc
A. Students taking SAT				
S-STATE (Equation 6)				
Squared bias (SB)	2.31	4.23	2.82	4.27
Pearson correlation	.42	.58	.32	.51
Marginal bias ($\phi - 1$)	.17	.25	.10	.26
(standard error)	(.03)	(.05)	(.03)	(.05)
B. Students NOT taking SAT				
S-STATE (Equation 6)				
Squared bias (SB)	.63	1.11	.65	.95
Pearson correlation	.78	.44	.72	.55
Marginal bias ($\phi - 1$)	−.02	−.06	−.01	−.04
(standard error)	(.02)	(.03)	(.02)	(.03)

composition has varied over time, the influence of self-selection has been offered as an explanation of some aspects of time trends in the data (CBO 1986, 1987). Similarly, the composition of test takers differs by state, suggesting that state means are distorted by sample selection bias.

The importance of SAT-type selection is directly analyzed by creating synthetic cohorts of HSB students based upon their indication of whether they have taken the SAT.[18] The impact of selection effects is shown in Table 4, which presents summary statistics for the fully specified state model [Equation (6)] estimated both for the sample who took the SAT and for the sample that did not take the SAT.[19]

The biases introduced by sampling only students taking the SATs is most pronounced. Differences in the quality of schools in the states sim-

18. The aggregate pattern of test taking reported in HSB is consistent with national patterns: The correlation between the fraction of the students in HSB who report having taken the SAT test, by state, and the publicly available figures on participation rates for the SAT is .94. This section concentrates on the SAT because it can be directly compared with other analyses and because ACT data are not generally available.

The aggregate correspondence of HSB data and observed national data lends further support to the approach in this paper of investigating specification problems with this consistent data base.

19. The alternative specifications were also estimated for the split sample but provided no additional insights. In general, the squared bias increased, the correlations with the true coefficients decreased, and the marginal bias increased slightly.

ply become confused with individual differences in ability, and there is a general upward bias in estimated school effects with low correlations and high squared bias. The marginal bias in school estimates from this selective sampling is roughly the same as that for ignoring demographic factors in the full sample (*S-PRETEST*), but the squared bias is much larger and the correlations with true estimates much lower. The misestimates for the sample not taking the test are noticeably less, perhaps reflecting the larger and more random sample of students.

E. Adjustments to Published SAT/ACT Data

Various attempts have been made to deal with the self-selection problems present in published SAT data. Maybe the most straightforward is the approach of the U.S. Department of Education, which only reports SAT or ACT data for states with a substantial proportion of test takers. Powell and Steelman (1984) and Dynarski (1987) as an alternative pursue statistical corrections for bias in the aggregate SAT data.

Figure 1 shows the obvious negative correlation between SAT scores and test taking percentages across states. The "wall charts"[20] of the U.S. Department of Education simply eliminate all observations with fewer than 31.5 percent test takers—the vertical line in Figure 1. While this procedure eliminates the sample correlation between scores and test taking patterns,[21] such arbitrary truncation cannot adequately deal with selection effects which almost certainly will be continuous.[22] In order to assess quality differences across states, the wall charts then add ACT tests to the remaining states, reflecting the fact that states with low participation in SATs have high participation in ACTs. Thus, they merge two truncated samples.[23]

20. See "State Education Statistics: Student Performance, Resource Inputs, and Population Characteristics, 1982 and 1985," prepared by the U.S. Department of Education, Office of Planning, Budget and Evaluation, Planning and Evaluation Service.

21. The correlation goes from −.86 for all 50 states to .09 for the 21 states left after truncation.

22. A general discussion of truncation in a variety of contexts can be found in Maddala (1983).

23. Dealing with varying participation rates by selecting the dominant test in a given state causes serious evaluation problems. The tests have a decidedly different geographical pattern (see Figure 1 where census regions of truncated and remaining states are given). The ACT is essentially a test for those west of the Mississippi, and the SAT a test for those east of the Mississippi. The use of different tests is likely to obscure any factors that vary systematically by regions. Further, they have distinctly different designs and interpretations [see CBO (1986)], and there is no generally accepted means for equating the scores on the two different tests.

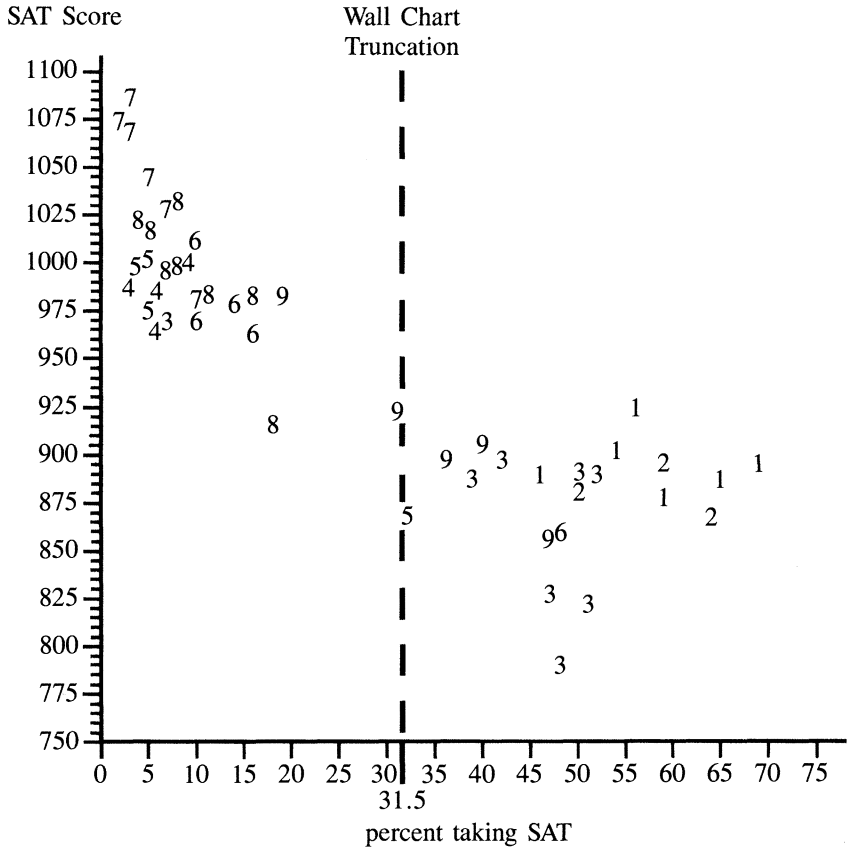

Figure 1
Relationship between State Average SAT Score and Percent of High School Graduates Taking the Test: 1982 by Census Region

Region key: 1=Northeast; 2=Middle Atlantic; 3=East North Central; 4=West North Central; 5=South Atlantic; 6=East South Central; 7=West South Central; 8=Mountain; 9=Pacific.

The top panel of Table 5 presents Pearson correlation coefficients comparing published SAT and ACT test scores by states to the estimates of base quality differences previously presented (*S-BASE*). While the truncation of the SAT sample improves the correlation, this evidence establishes rather conclusively that the aggregate SAT and ACT scores do not provide acceptable measures of school quality differences across the states. All but one coefficient is significantly different from one. These truncated aggregates of course suffer from the same problems discussed

Table 5
Comparisons of State School Quality Based on Published SAT/ACT with Fully Specified Estimates (S-BASE) *(for states with greater than 25 individual observations)*

Test/Sample	Vocabulary	Math	Science	Math + Voc
A. Unadjusted State Scores				
SAT				
All states	.10	−.20	.12	−.14
Wall chart SAT states	.69	.16	.60	.37
ACT				
Wall chart ACT states	.50	.39	.82	.38
B. Statistically Adjusted Scores[a]				
SAT				
All states				
Adjustment A	.58	.29	.64	.35
Adjustment B	.23	.06	.44	.05

Note: a. Adjusted state estimates from Powell and Steelman (1984).
Adjustment A: residuals from regressing average total SAT scores, by state, on test taking percentages.
Adjustment B. residuals from regressing average total SAT scores, by state, on test taking percentages, percentages squared and demographic variables.

previously about raw HSB scores: they omit consideration of other influences on performance.

The alternative approach of Powell and Steelman (1984) is to correct the aggregate scores statistically for other factors. They compute school quality as the residual from a regression of state SAT scores on participation rates (a direct measure of selectivity), racial composition, sex composition, and median incomes. Thus it is quite similar to *S-DEMOG* in underlying specification. The bottom panel in Table 5 presents the correlations of their adjusted SAT scores with the base state measures. The adjustment just for participation rates yields estimates of school effects that are significantly correlated with *S-BASE,* but are also clearly not its statistical equivalent. The estimates which come from the more complex demographic adjustment fit the data from Figure 1 more closely but do not compare well with the school quality estimates here. Only for the science test are these Powell and Steelman corrected state quality estimates significantly correlated with the base model. The simple wall chart adjust-

ment of deleting states with low test representation actually leads to a closer match with the base model than does the more complex demographic adjustment of Powell and Steelman.[24]

As indicated previously, using information about pretest scores is much more powerful than demographic adjustments both for dealing with general measurement problems and for dealing with selection issues. This suggests that the alternative SAT adjustments of Dynarski (1987) are likely to bring us closer to base state differences in achievement. His analysis, which uses panel data of state level SAT performance over ten years, incorporates fixed effects (separate intercepts) for individual states. This is essentially equivalent to considering differences in individual growth. Unfortunately, it is not possible to compare directly the state school quality estimates developed here with the implications of Dynarski's analysis.

IV. Conclusions

This study is exploratory, attempting to analyze and to disentangle the various factors entering into state-by-state variations in school performance. While previously neglected, approximately 10 percent of the variation in school quality falls between states. With increasing pressures toward accountability and with desire to evaluate the effects of major policy differences, there is, however, an unfortunate tendency to grasp at any available output or performance data.

The primary motivation of this analysis was the development of adequate measures of school quality, but the findings provide a number of insights into general analytical issues. Not surprisingly, raw test score differences across states are very misleading indices of school quality. Indeed, our work shows that mathematics tests, which enjoy a certain popularity because of their perceived objectivity, are particularly susceptible to bias from misspecification and sample nonrandomness. Moreover, while standardizing for differences among families improves estimates of state school quality, substantial error remains. The error confuses the ranking of states and, in any subsequent work, makes analyses of educational policies difficult to interpret. The best information for estimating school quality differences is value-added data, or data on growth over time in achievement of students. Moreover, state level value-added adjustments do remarkably well. The results also support the commonly held view that biases from raw test scores *or* from adjusted scores based

24. The Powell and Steelman adjustments employ sex, race, and income—exactly like the demographic adjustments in this paper.

on commonly available demographic characteristics work to overstate school influences.

The problems with specialized scores such as those on the SAT and ACT tests are even worse. Selection effects that evolve through differential participation rates in scores have important impacts on estimated state performance. Further, these selection effects cannot be adequately handled through simple devices such as deleting states with very low participation rates or introducing a participation variable into the analysis. Demographic adjustments to state differences also will not eliminate the biases.

The estimates of biases in measures of school quality from various misspecifications almost certainly are also relevant to questions of school-by-school comparisons. Indeed, while not analyzed here, the problems at the school level might be more severe because of the nature of school choice and the correlations thus generated.

Appendix 1

In the interests of student privacy, the compilers of High School and Beyond have declined to make available any information on the states in which the schools are located (if in fact such information exists). In order to make comparisons across states, therefore, the following algorithm has been used to identify, where possible, the probable state of residence.

Information on the student's post-secondary schooling is included in the HSB data. According to the survey of *Residence and Migration of College Students, 1975* by the National Center for Education Statistics, students tend to consume their post-secondary schooling close to home. Therefore, if a large percentage of the students from a particular school attended post-secondary institutions in a given state, then it is reasonable to conclude that their high school is also located in that state. States of residence are assigned to high schools on the basis of this argument.

Of course, there are students and institutions which confound this principle. It would be inappropriate to try to identify those students likely to want to move farther from home, but it is a relatively straightforward matter to identify those institutions most likely to attract students on a national rather than a local basis. For these purposes, an institution is defined as attracting students on a national basis if more than 60 percent of its freshman class is from out of state.[25]

25. Information on the fraction of students who are instate residents comes from *The College Board Handbook 1982–1983* published by the College Entrance Examination Board. The information was gathered by the College Board from material self-reported by the institutions on questionnaires sent to eligible institutions in November 1981. Eligible institu-

Combining the data on both the seniors and the sophomores by the school identification code (SCHLID) results in a distribution of the institutions at which students from each school received their post-secondary education, and a corresponding distribution of the states in which those institutions are located. Institutions for which there was no state identified or which were identified as national are excluded from the distribution. If at least 70 percent of the institutions in the distribution are located in a given state, and the census region for that state is the same as the region reported by HSB for the SCHLID, then that state is assigned to the SCHLID. Further, if at least 40 percent of the institutions in the distribution are located in a given state, and at least 75 percent of those institutions which are located in the census region reported for the SCHLID are located in that state, then that state is assigned to the SCHLID. Schools which remain unattributed are assigned to state "XX," and excluded from consideration here. Schools identified by fewer than four observations are also excluded. For obvious geographic reasons, no school was assigned to the District of Columbia.

References

Aitken, M., and N. Longford. 1986. "Statistical Modeling Issues in School Effectiveness Studies." *Journal of the Royal Statistical Society, A* (149, pt. 1):1–26.

Boardman, Anthony E., and Richard J. Murnane. 1979. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* 52(2):113–21.

Congressional Budget Office. 1986. *Trends in Educational Achievement.* Washington, D.C.: GPO.

———. 1987. *Educational Achievement: Explanations and Implications of Recent Trends.* Washington, D.C.: GPO.

Dynarski, Mark. 1987. "The Scholastic Aptitude Test: Participation and Performance." *Economics of Education Review* 6(3):263–74.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in the Public Schools." *The Journal of Economic Literature* 24(3):1141–77.

Koretz, Daniel. 1987. "A National Report Card: Risks in the Comparative Use of Current Indicators." In *A National Report Card: The Promise and Perils of Comparative Indicators,* Chair. Arthur Wise. Symposium conducted at the

tions are those institutions which are listed in the *Educational Directory, Colleges and Universities, 1981–1982* published by the National Center for Education Statistics, and which offer some undergraduate programs. The response rate or accuracy of institutions on this question is unknown.

annual meeting of the American Educational Research Association, Washington, D.C.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Powell, Brian, and Lala Carr Steelman. 1984. "Variations in State SAT Performance: Meaningful or Misleading?" *Harvard Educational Review* 54(4): 389–412.

Theil, Henri. 1971. *Principles of Econometrics*. New York: John Wiley and Sons.

Wirtz, Willard et al. 1977. *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline*. New York: College Entrance Examination Board.