Drawing Meaningful Trends from the SAT

Andrew J. Coulson

March 10, 2014

Cato Institute, 1000 Massachusetts Avenue N.W., Washington, D.C. 20001

Drawing Meaningful Trends from the SAT

by Andrew J. Coulson

Abstract

Though measures of long-term academic performance trends are valuable to education policy analysts and policymakers, they have been hard to come by at the state level, where most education policy is made. Such data are either nonexistent prior to 1990 or are unrepresentative of statewide student populations. A continuous series of state mean SAT scores is available back to 1972, however, and a previous study identified a method for adjusting these scores to account for varying participation rates and demographics. The present paper aims to extend that earlier work, modifying it to facilitate meaningful trend analysis, using a substantially larger data set, relaxing its assumptions, and improving its approach to model selection and validation. The optimal model derived from that process is externally validated against both lagged 8PPth-grade and contemporaneous 12th-grade National Assessment of Education Progress (NAEP) scores, with good results.[i]

***

Andrew J. Coulson is director of the Cato Institute's Center for Educational Freedom and author of *Market Education: The Unknown History*.

[i] I would like to thank Mark Dynarski, Philip Gleason, Eric A. Hanushek, Patrick Wolf, and two anonymous reviewers for helpful comments on earlier drafts of this paper. Any remaining errors are my own.

Introduction

Though measures of long-term academic performance trends are valuable to education policy analysts and policymakers, they have been hard to come by at the state level, where most education policy is made. True to its name, the Long Term Trends (LTT) series of the National Assessment of Educational Progress (NAEP) is a consistent and nationally representative longitudinal metric reaching back over 40 years. But LTT results are not available by state. A separate suite of tests, the Main NAEP, does report state-level results, but these only begin in the 1990s and the content of the tests changes over time, making it arguably less useful as a longitudinal measure.

Results for college entrance tests such as the ACT and the SAT are also available at the state level and have two advantages over the Main NAEP: they measure performance near the end of high school (giving a more complete picture of any K–12 educational impact), and they reach back many decades. Unfortunately, the American College Testing program does not divulge state ACT results prior to 1992, on the grounds that they are unreliable. The College Board, by contrast, releases state SAT data back to 1972.

Still, these tests have a shortcoming of their own for the purposes of state trend analysis: they are taken by self-selected subgroups of students that are not representative of their peers statewide. Despite this problem, the media and policymakers sometimes rely on raw SAT scores as an academic performance trend metric (e.g., Layton and Brown, 2012)—a dubious practice.

In the 1980s, numerous researchers proposed addressing the SAT's shortcomings as a measure of state performance by adjusting its scores based on participation rates. A key weakness of these efforts was that there was no basis for external validation—no separate source of data on the relative performance of the states against which the adjusted SAT scores could be evaluated. Not surprisingly, the adjusted scores offered by these models differed from one another, and there was no way to know which, if any, were to be favored (Dynarski and Gleason, 1993; hereinafter D&G).

That changed with the publication of the first 8th grade Main NAEP test scores for 38 states, drawn from representative samples of each state's population in 1990. D&G used these NAEP scores to validate an SAT adjustment model. They began by developing an education production function to predict SAT scores using test-participation rates as well as state population and school-system characteristics, fitting 20 years' worth of observations (1971–1990) using a fixed effects panel regression. D&G then produced adjusted predicted SAT scores from this model for the year 1990, setting the participation rate for every state equal to the overall mean participation rate—thus controlling for the key difference between the SAT and the NAEP. Finally, they ranked their adjusted SAT scores for the 38 states that also participated in the NAEP, and compared those rankings to the NAEP rankings. The correlation coefficient for this comparison was 0.78.

D&G noted three potential weaknesses with this procedure. First, at the time, the Main NAEP tested only 8th graders, whereas the SAT is taken chiefly by 12th (and to a lesser extent 11th) graders. Second, only public-school students were tested for the 1990 NAEP, whereas both

public- and private-school students take the SAT. And, third, the 1990 NAEP covered only mathematics, whereas the SAT covered both verbal and mathematics skills. The NAEP would thus have been a poor validation metric for adjusted SAT scores if: 1) there were a low correlation between the scores of 8th graders and those of older students; 2) there were a low correlation between the scores of public and private-school students within a given state, or if there were a substantial public/private performance gap and private school enrollment rates varied substantially between states; or, 3) if there were a low correlation between students' verbal and mathematical performance.

D&G believed, but could not prove, that these problematic conditions were unlikely to obtain, and their results depended on these assumptions. In the present paper, no such assumptions are necessary. Limited state-level NAEP data for the 12th grade are now available, and their degree of correlation with 8th-grade scores can be directly measured. More recent NAEP data also include both public- and private-school students, and so this difference with the SAT test takers has been eliminated. Finally, recent NAEP data include both reading and mathematics scores, and so we can compare adjusted composite SAT scores to combined NAEP mathematics and reading results.[1]

Several additional improvements and extensions to the D&G procedure are also now possible, and these are discussed in the methodology section. The original D&G model holds up to an impressive degree when applied to a newly expanded dataset. Nevertheless, the model developed here noticeably improves upon its results and does so with a substantially higher degree of confidence thanks to the use of independent validation data.

Using the model developed in this paper, the author will, in a future publication, estimate 40-year trends in average state academic performance over time, controlling for demographic factors and SAT participation rate. This will offer education analysts and policymakers a new tool for evaluating the policies of the past and informing future decisions.

Data, Empirical Model, and Methods

The SAT Data

The test data for this study were state average SAT scores for the years 1972 through 2012, provided by the College Board. Because the score scale of the SAT was recalibrated ("recentered") in 1995, the raw scores before and after that year are not directly comparable. However, the College Board has a conversion formula for placing the pre-1995 scores on the new re-centered scale (Dorans 2002), and it is this fully recentered data series that was used in the present study, to ensure consistency over time. Though the conversion formula does not make the pre- and post-recentering scores perfectly interchangeable, the remaining deviations are slight, having little impact on overall state averages.

The Dynarski and Gleason Approach

The functional relationship between SAT participation rate and test scores is not known a priori; nor are student-level data available to facilitate the estimation of that relationship. D&G addressed this challenge by evaluating several different regression models. Following Hanushek (1979), they set out to estimate the following set of education production functions:

$$SAT_{it} = X_{it}\beta + \alpha f_n(p_{it}) + \gamma_i + \delta_t + \varepsilon_{it},$$

where SAT_{it} is the average SAT score for state i in year t, X_{it} is a vector of characteristics of state i in year t, $f_n(p_{it})$ is one of n alternative functions of the participation rate, γ_i is the fixed effect for state i, δ_t is the fixed effect for year t, and ε_{it} is a random-error term.

D&G regress this set of models using 20 years of data, from 1971–1990. Refining earlier attempts at adjusting SAT scores for participation, they define the participation rate as the number of SAT test takers divided by the state's population of 18-year-olds, rather than dividing by the state's grade-12 enrollment. This, the authors hoped, would eliminate a potential source of error due to variations in dropout rates across states (which would alter the meaning of a participation rate based on 12th-grade enrollment).

D&G consider seven variations on this model: a linear function of the participation rate without control variables, as well as linear, cubic, and logistic models that include a common set of control variables and alternately include or omit fixed effects. A concern with this approach is that we do not know, in advance, if the relationship between state SAT scores and the participation rate is linear, cubic, logistic, or something else altogether. And, to the extent that relationship is misspecified, it will bias the model's predicted SAT scores.

D&G do offer an empirical method for choosing among their various models. For each model, they generate predicted state SAT math scores for 1990, while holding constant the SAT participation rate at the national mean. They then compute the correlation between the state ranking that results from their adjusted SAT scores and the state ranking on the 8th grade NAEP mathematics test administered in the same year. Since only 38 states participated in the NAEP in that year, D&G are limited to ranking only those states.

Because the NAEP is given to representative samples of students rather than to self-selected subgroups, it is a reasonable benchmark for assessing the plausibility of SAT score-adjustment models. Using that benchmark, D&G identify the following as their preferred model:

$$\begin{aligned} SAT_{it} = {} & \text{poverty}\times\beta_1 + \text{household\_size}\times\beta_2 + \text{pct\_white}\times\beta_3 + \text{tchr\_salary}\times\beta_4 + \\ & \text{spending}\times\beta_5 + \text{tchr\_student\_ratio}\times\beta_5 + \text{pct\_public\_enrollment}\times\beta_5 + \\ & \log(p_{it}/1-p_{it}) + \gamma_i + \delta_t + \varepsilon_{it}. \end{aligned}$$

The control variables are household size, percent of state population that is white, percentage of the population living in poverty, average salary of instructional staff in public schools, per pupil expenditures in inflation-adjusted dollars, the teacher/student ratio, and the percentage of the secondary student population enrolled in public schools.

D&G were not able to independently validate this model, however, since they were obliged to use the only available year of external NAEP data in selecting their preferred model. There were thus no independent NAEP data left with which to validate the selected model.

Improvements and Extensions to Dynarski and Gleason

The D&G methodology described above can be improved upon in several ways. First, as a peer reviewer of the present paper suggested, we can compute correlation coefficients between the NAEP and adjusted SAT scores directly, rather than for the state rankings corresponding to those scores. It is not clear why D&G chose to use ranks, but doing so reduces the amount of

information available for the correlation and thereby renders the resulting coefficients less informative.

Second, combined SAT reading and mathematics results can now be compared to the combined reading and mathematics NAEP results, since both NAEP subjects are now available for all states over several years. The raw scores from the NAEP reading and math tests cannot be directly averaged, however, since they have different standard deviations. To account for this, the raw NAEP scores for each subject are first standardized to a mean of 0 and a standard deviation of 1, producing what are known as z-scores, and then those z-scores for each subject are averaged together. That average can then be compared to the predicted composite SAT score for the corresponding state and year.

A third improvement made possible by the expanded set of NAEP data now available is to introduce a four-year lag between the NAEP and SAT scores to be compared. The highest grade for which state NAEP scores are available is the 8th grade, while the SAT is administered primarily to 12th graders. So, by lagging the comparison by four years, we can maximize the extent to which scores for the same cohort of students are being compared.

Each of the above improvements could be implemented within the same methodology used by D&G, but that methodology can itself also be extended. The first such extension is to relax D&G's assumption that the relationship between participation rate and state average SAT scores is necessarily either linear, cubic, or logistic. Instead, the present paper evaluates 1,023 different possible functions of the participation rate. These alternative functions correspond to every possible combination of between one and five of the following transformations of participation rate, p:

$$log(p),\ log(p/1\text{-}p),\ p^{0.1},\ p^{0.3},\ p^{0.5},\ p,\ p^{2},\ p^{3},\ p^{4},\ p^{5},\ p^{6}.$$

In other words, we wish to consider the predictive power of each of these terms individually, as well as every possible combination of two, three, four, and five of these terms. The total number of such combinations is 1,023. Since that set encompasses the three functions of participation rate explored by D&G, the present paper makes it possible to empirically test their models against a wide range of alternative specifications.

A second extension to D&G is to relax their assumption that all of the control variables have a linear relationship to state mean SAT scores, as there is no obvious justification for that assumption. Some of the controls could, for instance, be related to SAT scores quadratically or logarithmically, rather than linearly. Failing to test for these alternative transformations of the control variables could lead to a misspecification of the equation and correspondingly biased results.

Another unstated assumption in the D&G model, equally without theoretical or empirical justification, is that there are no interactions among the control variables. If such interactions exist, a model omitting them will again be misspecified and its results biased.

Finally, several variables that may be related to state average SAT scores, such as the mean educational attainment of the state's adult population, are not included in the D&G model, potentially introducing omitted variable bias.

To mitigate these potential misspecification problems, the present study considers a larger list of possible controls, including nonlinear transformations and interaction terms. That master list of control variables appears in Table 1.

Table 1. Master List of Control Variables

grd12pct	ratio of 12th-grade student population to 18-year-old population
grd12pctsq	*grd12pct* squared
Pctwhite	fraction of the population that is white (all ethnicities)
Pctwhitesq	*pctwhite* squared
Hispenroll	percentage of students who are Hispanic (all races)
Hispenrollsq	*hispenroll* squared
Blackenroll	percentage of students who are African American (all ethnicities)
Blackenrollsq	*blackenroll* squared
Pcths	fraction of the population who have at least a high school diploma
pcthssq	*pcths* squared
pctba	fraction of the population who have at least a bachelor's degree
pctbasq	*pctba* squared
pctnative	fraction of the population that is Native American or Alaska native
pctnativesq	*pctnative* squared
pctnative5	*pctnative*^5
pcthwn	fraction of the population that is Hawaiian or other Pacific Islander
pcthwnsq	*pcthwn* squared
pcthwnln	natural log of *pcthwn*
pctpublic	fraction of students who attend public schools
pctpublicln	natural log of *pctpublic*
pctpublicsq	*pctpublic* squared
ppspend	public school per pupil spending, constant dollars
income	median state income
incomesq	*income* squared
urban	fraction of population living in urban areas (std'zd to mean=0, sd=1)
income_urb	interaction term: *income* × *urbanicity*
income_grd12	interaction term: *income* × *grd12pct*
ba_public	interaction term: *pctba* × *pctpublic*
ba_grd12	interaction term: *pctba* × *grd12pct*
ba_urban	interaction term: *pctba* × *urbanicity*

Note that some demographic variables in Table 1 pertain to the state population as a whole while others pertain specifically to the student population, due to the particular data sets that were readily available at the time this study was conducted.

As it happens, the enhancements described above create concerns of their own. The first and most serious is overfitting, which occurs when a model fits not only the underlying relationship of interest but also the random errors specific to a particular dataset. The result a model that is misspecified and, on new datasets, produces a poor fit (i.e., exhibits poor predictive validity). The risk of overfitting increases with the number of variables in the model and the number of different models examined, and so must be addressed.

Another problem associated with having large numbers of control variables is multicollinearity, which arises when two or more of the control variables are correlated with one

another, or with linear combinations of each other. This means that the controls are, to some extent, duplicative in attempting to explain the same variations in the underlying data. Multicollinearity drives up the standard errors of the affected variables, reducing their statistical significance, and potentially rendering otherwise significant variables insignificant.

For both of these reasons, it is advantageous to find the most parsimonious (smallest) model that adequately explains the data. A common technique for doing this is principal component analysis (PCA), which uses linear algebra to transform a large (and potentially collinear) set of control variables into a smaller set of linearly uncorrelated vectors that more succinctly explain the bulk of the variation in the dependent variable. As discussed in Appendix A, however, PCA is not the ideal approach to the current problem.

The alternative procedure adopted in this paper is an extension of the one adopted by D&G, in which the various models under consideration are rated based on the correlation between their predicted SAT scores and independent state NAEP data. Then, as a hedge against overfitting, we validate the preferred model using additional independent NAEP data that were not used during model selection.

The challenge with this approach in the present context is that while D&G only compared seven different models, this study seeks to evaluate all the parsimonious models that could possibly be drawn from a set of 41 independent variables (11 different transformations of the participation rate combined with 30 controls). The number of different combinations of, say, 14 variables drawn from this set of 41 would be 3.5e10—an impractical tally.

An initial concession to practicality is to separate the selection of participation-rate functions from the selection of control variables. This eliminates the combinatorial problem in the case of the participation rate function, since, as noted above, the number of combinations of between 1 and 5 terms drawn from a set of 11 terms is only 1,023—a tractable number.

In the case of the control variables, choosing 14 variables from a list of 30 would still yield a prohibitive number of combinations. A further concession to practicality is to allow experience with the education research literature to guide the selection of control variables based on those that are often found to be significant predictors of academic performance. This is a common heuristic solution to a frequent problem in model design. By winnowing down the list of 30 possible control variables to just 21, we can tame the combinatorial problem, since 21-choose-14 produces only 116,280 combinations—a number amenable to contemporary statistical software and computer hardware. In other words, we can feasibly execute 116,280 regressions, use them to generate predicted SAT scores, and then test those predictions against corresponding NAEP data in the manner described above, picking the best among them.

Rather than be content with a single winnowed-down set of 21 controls, however, we can explore the controls listed in Table 1 more fully by repeating the above control selection process based on the results of each previous iteration. The best predictors from our initial subset of 21 can be kept, and the weakest replaced with others from the original set of 30 that have not yet been tried.

That leaves us with one further question: how can we jointly determine both the preferred function of participation rate and the preferred set of control variables, given the combinatorial impossibility of identifying them simultaneously? An obvious solution is to approach the two

problems sequentially. First, we can search for a preferred set of control variables assuming some plausible function of participation rate. Next, we can search for a potentially superior function of the participation rate while assuming the set of controls identified in the previous step. Then we can return to the first step, using the new participation rate function in place of the initial one. This iterative algorithm—known as successive approximation—can be repeated until the overall results of the combined model no longer improve, as measured by our NAEP correlation test. This procedure is laid out in more detail below.

1) Choose an initial participation rate function for use in the first iteration of our procedure. We take D&G's preferred function of participation rate, $\log(p_{it}/1-p_{it})$, as a promising starting point.

2) Combine the given function of participation rate with all 116,280 different sets of 14 control variables chosen from a sublist of 21 (itself drawn from the 30 controls in Table 1, based on knowledge of the education research literature). Run all the 116,280 resulting time series regressions. For each of those regressions, predict the adjusted SAT scores for 1996, 2007, and 2011, setting the participation rate to the national mean. Next, correlate those adjusted SAT scores with the composite 8th-grade NAEP z-scores from 4 years earlier.[2] Select as the preferred combination of controls the one whose predicted SAT scores maximize the average of the three correlation coefficients.

3) Select a new SAT participation rate function by regressing SAT scores against the set of controls just identified in Step 2, paired in turn with each of the 1,023 different functions of participation rate discussed above. For each of those regressions, predict the adjusted SAT scores for 1996, 2007, and 2011, setting the participation rate to the national mean. Next, correlate those adjusted SAT scores with the composite 8th-grade NAEP z-scores from 4 years earlier. Select as the preferred function of participation rate the one whose predicted SAT scores maximize the average of the three correlation coefficients.

4) Using the participation rate function just identified in Step 3, start over at Step 2. Repeat this process until the NAEP correlation results of the preferred model cease to noticeably improve.

5) As a hedge against overfitting, externally validate the preferred model by computing the correlation coefficient between its adjusted predicted SAT scores for 2009 and the NAEP 8th-grade scores for 2005 (which have not been used during model selection). As a frame of reference, compare that correlation value with the correlation between NAEP 8th-grade test scores for 2005 and NAEP 12th-grade test scores for 2009 (available for 11 states). Finally, directly validate the adjusted SAT scores for 2009 against the NAEP 12th-grade scores for 2009.

Though computationally intensive and somewhat ungainly, this iterative approach has the advantage of producing results that are substantially superior to those of D&G as well as to those of principal component analysis, as discussed in the Results section and in Appendix A.

Other Model Differences from Dynarski and Gleason

Prior to moving on to the results, two further deviations from the D&G approach are required due to the somewhat different goal of this paper. As noted in the introduction, the goal here is to identify meaningful trends in student academic achievement over time at the state level. We are looking for any changes in performance that can be reasonably attributed to changes in the quality of academic instruction. Is instruction improving, declining, or staying the same? To answer that question, we must control for any factors that may affect average state

SAT scores but that are unrelated to changes in the quality of instruction. That is why we control for the student-participation rate and for demographic characteristics of students and parents.

To the extent that we successfully control for all noninstructional factors, the variation over time between a state's actual SAT scores and our model's predicted SAT scores can plausibly be attributed to variation in the quality of instruction being offered. The differences between actual and predicted values in a regression are known as the residuals, and so our goal is to push the effects of all instructional factors into the residuals, and pull everything else out into control variables. This will allow us, in a subsequent paper, to chart the trends in the residuals as an estimate of trends in instructional quality over time.

A corollary of that goal is that our model cannot include as control variables any education-policy factors that can reasonably be expected to affect instructional quality. If we controlled for such factors, it would pull their variation out of the residuals, and so would not show up in our trend charts. For instance, imagine that teacher salaries are related to student achievement, and that the state of Washington gradually increased the portion of its budget allocated to teacher salaries over time. If we were to control for teacher salaries in our regression model, the improvements in SAT scores that resulted would be pulled out of the residuals of the model and captured by the teacher-salary variable. As a result, the residuals trend chart for Washington would not show the real improvement in instructional quality that took place in that state. Due to the need to keep the influence of such policy factors in the residuals, the control variable list in Table 1 excludes both teacher salaries and the student/teacher ratio.[3]

A peer reviewer rightly observes, however, that if these deliberately omitted control variables are highly correlated with another regressor (or combination of regressors) in the final model, then it hasn't really been omitted at all. If that's the case, the usefulness of the model's residuals as a gauge of school effects on academic performance would be partially compromised. To test for simple collinearities we can separately check the correlations between the two omitted variables and each of the controls in the final model; and to test for multicollinearity we can compute the variance inflation factors for the omitted variables. These tests are performed in the following section.

It might be argued for the reasons just discussed that per pupil spending should also be excluded from the model, but this is arguably a somewhat different case since neither enrollment nor the total funding available for public schools are policy decisions under the control of public-school officials (the latter being determined by voters and their elected legislators). Hence, it seems reasonable not to exclude per pupil spending from the model. At any rate, the question is academic because per pupil spending does not make it through the empirical model selection process described above, failing to appear in the preferred model.

This argument for excluding policy variables does apply, however, to *year* (but not state) fixed effects. Consider, for instance, the possibility that some federal education-policy initiative is introduced in a given year, affecting every state, and that this initiative improves student outcomes. If year fixed effects are included in the model, they will capture any such uniform improvement in student performance over time resulting from a new federal education policy, pulling that improvement out of the residuals. Trend charts of the residuals would then no longer show the improvement that had taken place. For this reason, year fixed effects must be excluded from our model.

Note that D&G did not need to exclude these variables from their own model because they were not trying to measure *trends*, but rather to compare states to one another at one particular point in time.

Results

The preferred control variables identified by the above iterative process are presented in Table 2, along with their descriptive statistics. In this section, variable descriptions are used in place of abbreviated variable names, to improve readability.

Table 2. Descriptive Statistics of Preferred Controls

Variable Description	Mean	Standard Deviation	Minimum	Maximum
% of population that is white	0.8561	0.1138	0.3302	0.9956
% of population with at least a BA degree	20.72	6.214	7.489	39.34
% of students in public schools	0.8973	0.0511	0.7055	0.9876
(% of students in public schools)^2	0.8078	0.0902	0.4977	0.9754
(% of pop'n with at least a HS diploma)^2	5962	1611	1659	8612
median state income × urbanicity	2684	50927	-147649	121199
% of students who are Hawaiian	0.4999	2.716	0.0039	26.72
log(% of students who are Hawaiian)	-2.706	1.417	-5.545	3.286
(% of students who are Hispanic)^2	183.0	479.7	0.0000	3815
% with BA × % students in public schools	18.57	5.563	7.060	35.85

One notable difference from the D&G model is that neither *income* (standing in for D&G's poverty measure) nor *per pupil spending* make it into the preferred model, apparently conferring little added ability to predict NAEP scores (though income does feature indirectly through the interaction term *income × urbanicity*). Interestingly, *per pupil spending* is nevertheless a statistically significant predictor of SAT scores, and, as in D&G's model, its coefficient is negative (β= -.0016, SE=.0004, $p < 0.001$).

A few words are perhaps in order on the model's inclusion of the *log of the percentage of students who are Hawaiian*. This regressor suggests that changes at the very low end of the range of the Hawaiian population are associated with larger shifts in the SAT score than changes at the high end of its range. Intuitively, it seems highly unlikely that variations in low levels of Hawaiian population have a significant direct effect on state mean SAT scores. Rather, the significance of this term is likely due to collinearity with an unknown omitted variable. Since we are not using the coefficients to make judgments about the SAT performance of Hawaiian students, this probable collinearity does not diminish the value of this term in our model. Our goal is only to produce the best possible estimates of adjusted SAT scores, not to apportion the explanation for scores among the control variables.

Following D&G and the broader time series regression literature, we perform a Hausman test comparing the fixed and random effects versions of the preferred model. This test rejects the null hypothesis that the coefficients of the two versions are the same at the $p < 0.001$ level, and hence the fixed effects model is to be preferred—as was the case in the D&G paper.

Next, we can investigate possible correlations between the two deliberately omitted control variables (*teachers' salaries* and *pupils per teacher*) and the regressors in the final model. As shown in Table 3, the highest correlation between the omitted variables and the final model controls is 0.52, and most are substantially lower. The variance inflation factors for the two omitted variables are 4.7 and 4.9, respectively, which are below the cutoff value of 10, at which multicollinearity is commonly considered problematic.

Table 3. Omitted Variable Correlations with Model Regressors

Regressor	Teacher Salary (tchr_salary)	Pupils per Teacher (pupils_p_tchr)
participation rate function	-0.49	0.12
% of population that is white	-0.13	-0.06
% of population with at least a BA degree	0.31	-0.39
% of students in public schools	-0.34	0.13
(% of students in public schools)^2	-0.34	0.14
(% of population with at least a high school diploma)^2	0.15	-0.39
median state income × urbanicity	0.52	0.24
% of students who are Hawaiian	0.08	0.09
log(% of students who are Hawaiian)	0.19	0.07
(% of students who are Hispanic)^2	0.05	0.11
% with BA × % students in public schools	0.24	-0.36

After the repeated iterations of Steps 2 and 3, we are left with a slightly different function of participation rate than the one preferred by D&G:

$$f_2(p) = \alpha_1 \times p^{0.5} + \alpha_2 \times p^4 + \alpha_3 \times p^5$$

The results of combining this model with the optimal parsimonious control variable set described above are presented in Table 4. Its average NAEP-to-adjusted-SAT correlation coefficient is 0.85, and its adjusted r-squared value is 0.94.[4]

The *(% of population with at least HS diploma)^2* and *(% of students who are Hispanic)^2* coefficients are statistically insignificant at any conventional level, but their inclusion in the model nevertheless improves its ability to predict NAEP scores, which is the reason this model emerged from the iterative selection process as the preferred one.

Table 4. Preferred Model

Adjusted R-squared = 0.94 Root MSE = 16.5

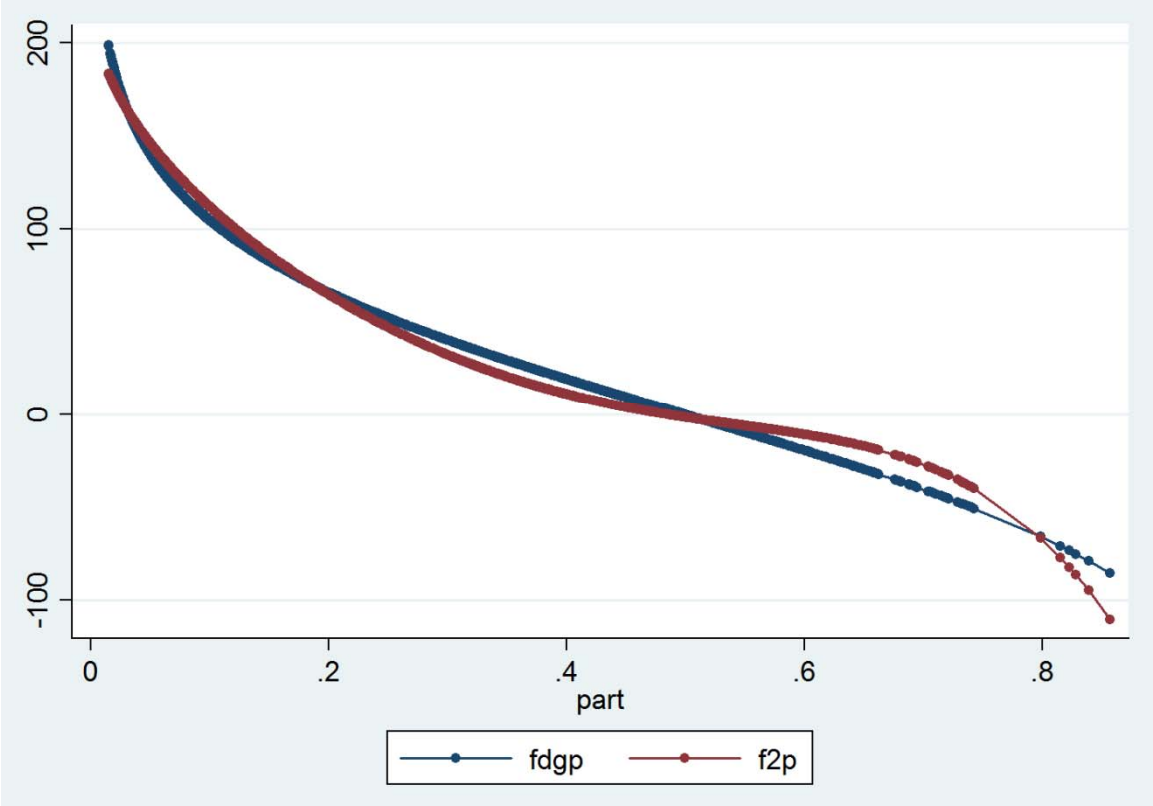
Regressor	Coefficient	Standard Error	t	P>t
$f_2(p)$	1.000	0.0265	37.74	0.0000
% of population that is white	280.553	38.53	7.280	0.0000
% of population with at least a BA degree	-5.781	1.842	-3.140	0.0020
% of students in public schools	810.193	414.825	1.950	0.0510
(% of students in public schools)^2	-550.856	239.749	-2.300	0.0220
(% of pop'n with at least HS diploma)^2	0.0012	0.0010	1.260	0.2080
median state income × urbanicity	0.0002	0.0000	3.740	0.0000
% of students who are Hawaiian	-1.874	0.6236	-3.010	0.0030
log(% of students who are Hawaiian)	16.14	1.484	10.88	0.0000
(% of students who are Hispanic)^2	-0.0021	0.0023	-0.9200	0.3560
% with BA × % students in public schools	7.291	2.067	3.530	0.0000

Interestingly, over most of the range of the participation rate variable, p, the value of $f_2(p)$ is very similar to that of D&G's preferred logistic function:

$$f_{d\&g}(p) = \alpha \times \log(p / 1 - p)$$

The two are charted together in Figure 1, showing that D&G's function holds up remarkably well despite the addition of 21 years of new observations. The only substantial departure between these functions occurs at the highest values of p. There is a likely explanation for this: during the years for which D&G had data, the highest state SAT participation rate was 0.62, and there were only three years in which the highest rate exceeded 0.6. However, in the years since 1991, the maximum participation rate has never fallen *below* 0.65. Moreover, after Maine adopted the SAT as a required statewide test in 2006, its participation rate jumped above 0.8. With no data to fit in this range, D&G had no empirical basis for modeling it.

Figure 1. Functions of SAT Participation Rate: Dynarski & Gleason vs. $f_2(p)$



Finally, turning to Step 5 of our procedure, the 2009 adjusted SAT scores produced by the preferred model can be externally validated against the 2005 average NAEP z-scores for 8th graders. The resulting correlation is 0.85, matching the average correlation obtained for the 1992×1996, 2003×2007, and 2007×2011 year pairs during model selection. It should be noted that adjusted SAT scores and NAEP scores varied during this time period, so the similarity between the correlations for the model selection and validation phases is not due to the underlying scores having remained constant. For comparison purposes, scatter plots of both raw and adjusted 2009 SAT z-scores versus 2005 NAEP 8th-grade z-scores are presented in Figure 2.

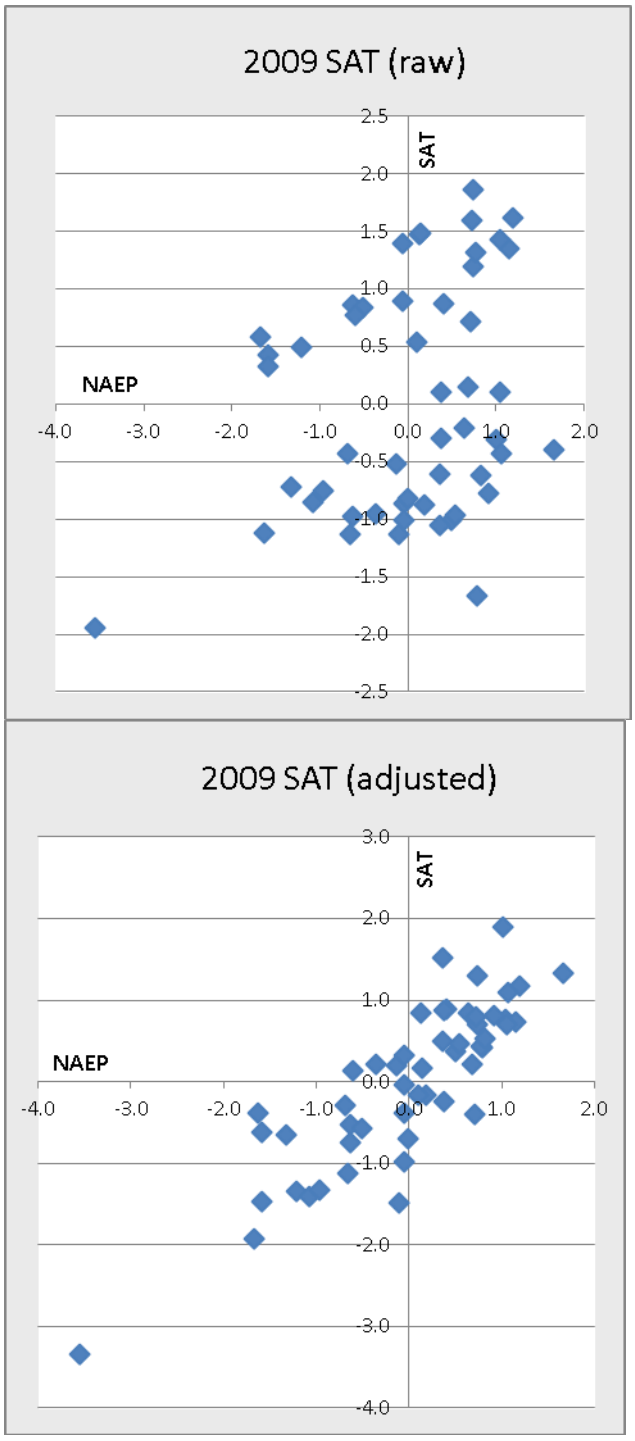
In assessing the results depicted in Figure 2, it must be remembered that even a perfect SAT adjustment model would yield a correlation coefficient below 1 when compared to lagged 8th grade NAEP scores. This is evident from the fact that the correlation between the 2009 12th-grade NAEP (available for 11 states) and the 2005 8th-grade NAEP is only 0.94. In that context, the preferred SAT adjustment model's NAEP correlation of 0.85 (across several different lagged year pairs) seems reasonable.

As further validation of the model presented in this paper, the correlation between the 2009 adjusted SAT and the 2009 12th-grade NAEP is 0.90, somewhat better than the results of the four-year lagged comparisons to 8th-grade students, as might be expected.

Note that even with a comparison of SAT to NAEP scores for the same year, a correlation coefficient of 1 is not to be expected. Nor, indeed, would it be expected even if the exact same test were administered to the same students just a few months apart. A given test is considered to

have good “repeatability” or “test-retest reliability” if the correlation coefficient of the results across two test administrations to the same students is equal to or greater than 0.8 (Drost, 2011). On the one hand, since we are comparing state mean scores rather than individual student scores, a higher degree of correlation should be expected. On the other, since we are comparing the results of two entirely different tests administered to two largely (perhaps even completely) different groups of students, we would expect a lower correlation. Considering these competing factors together, the preferred model’s same-year adjusted SAT to NAEP correlation of 0.90 seems reasonable.

Figure 2. 2009 SAT z-Scores vs. 2005 8th-Grade NAEP z-Scores
(measured in standard deviations, with mean of zero)



Some comparisons to the D&G results are also worth exploring. The present study's NAEP z-score validation correlations of 0.85 and 0.90 are superior to the 0.78 rank correlations reported by D&G, though their result was for 8th graders taking the NAEP in the same year as the (mostly

12th grade) SAT takers—thus representing entirely different cohorts of students. For comparison purposes we can repeat their test, computing the correlation between the 2011 adjusted SAT scores and the NAEP 8th-grade z-scores in that same year. The result, 0.84, is slightly lower than for the lagged correlations that match student cohorts, but still an improvement on the original D&G results.

Conclusion

The present study had two goals: an attempted replication of Dynarski & Gleason's state mean SAT score adjustment for student participation rates, using new data and a more rigorous model selection procedure, and an effort to draw useful longitudinal trend data from SAT scores.

To a remarkable degree, D&G's preferred model proved replicable (see Figure 2). The main deviation of their logistic participation rate function from the new empirical data of the past 20 years comes at the high end of the participation range—a range for which few data were available during the period (1972–1990) that they examined.

Using three separate sets of data for model fitting, model selection, and model validation, this paper yields correlations between adjusted SAT scores and lagged 8th-grade NAEP scores averaging 0.85. The correlation between adjusted 2009 SAT scores and NAEP 12th-grade scores for the same year is 0.90. These results compare favorably with D&G's correlation of 0.78 between SAT and NAEP rankings, particularly given that a correlation of 1 cannot be expected when comparing the results of different tests administered to different students.

In the absence of any state-level academic achievement metric with comparable scores reaching back to the early 1970s, the media and policy analysts sometimes cite unadjusted SAT scores as a measure of long-term achievement trends. This is inadvisable, given the known relationship between SAT scores and participation rates and demographics, both of which fluctuate not only between states but over time within states.

The present study makes it possible to control for these participation and demographic influences on SAT scores. Moreover, it modifies the D&G model to remove time fixed effects and independent variables that are under the control of public-school officials, in order to push any school effects on SAT scores into the residuals. As a result, trends in the residuals offer a rough estimate of any trends in school effects on SAT scores over time. Though not ideal, this is a better measure of such state-level effects than currently exists for the years prior to 1990, and so offers policy analysts and policymakers a new tool for evaluating past policies and informing future ones.

Appendix A

Principal Component Analysis Discussion

Principal component analysis treats an initial set of variables as a matrix of vectors, and uses linear algebra to transform them into an alternate set of orthogonal vectors (the principal components) that capture the variation in those initial variables more compactly. Original variables that are largely collinear are absorbed into a single new component. This eliminates the

problem of multicollinearity but turns out not to be the optimal tool for the task at hand in this paper.

When a PCA is performed on the control variables listed in Table 1, and the resulting principal components are used (along with a function of the participation rate) to produce adjusted 2009 SAT scores, the resulting SAT scores show an inferior correlation to the 2005 NAEP validation data. The PCA validation correlation is 0.76, versus 0.85 for the validation step reported by the preferred model presented in this paper.

It is perhaps not surprising that PCA yields noticeably inferior results on the NAEP validation, since PCA does not make use of the external NAEP data we have available to guide model selection. Principal component analysis looks only at the values of the control variables themselves in identifying the principal components that will comprise the final model. But, extending D&G's approach, the iterative process adopted in this paper takes advantage of state NAEP scores to help identify the best model. Indeed, with so many additional years of state NAEP scores now available, we use multiple years of NAEP scores to guide model selection instead of just a single year, while still being able to reserve additional independent years and grades of NAEP data as a final validation check on the selected model, to ensure that we have not fallen victim to overfitting.

References

Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger, "School Choice, School Quality and Postsecondary Attainment," National Bureau of Economic Research, NBER working paper 17438, TTUUhttp://www.nber.org/papers/w17438.

Dorans, Neil J., "The Recentering of SAT® Scales and Its Effects on Score Distributions and Score Interpretations," The College Board, Research Report No. 2002-11, http://www.ets.org/Media/Research/pdf/RR-02-04-Dorans.pdf.

Drost, Ellen A. "Validity and Reliability in Social Science Research." *Education Research and Perspectives* 38, no. 1 (June 2011):TT 105–23.

Dynarski, Mark and Philip Gleason. "Using Scholastic Aptitude Test Scores as Indicators of State Educational Performance." *Economics of Education Review* 12, no. 3 (1993): 203–11.

Hanushek, Eric. "Conceptual and Empirical Issues in the Estimation of Education Production Functions." *Human Resources* 14, no. 3 (1979): 352–88.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: an Experimental Evaluation." National Bureau of Economic Research, NBER working paper 14607, http://www.nber.org/papers/w14607.

Layton, Lyndsey, and Emma Brown. "SAT Reading Scores Hit a Four-decade Low." *Washington Post*, September 24, 2012, http://articles.washingtonpost.com/2012-09-24/local/35495510_1_scores-board-president-gaston-caperton-test-takers.

[1] NAEP scores cannot be directly averaged across subject, but a procedure for allowing them to be averaged is described in the next section.

[2] NAEP reading scores are not available for 1992, and so the NAEP math scores alone are correlated with the composite SAT scores in that year. As it happens, NAEP reading and math scores are typically correlated at about 0.94, so the math scores alone are a reasonable surrogate.

[3] Controlling only for exogenous variables that affect academic outcomes and then using the residuals to gauge teacher-level, school-level, or state-level education effects is not uncommon in the education production function literature (see, e.g., Kane and Staiger, 2008; Deming et al., 2011).

[4] Note that these are the results with data for the District of Columbia excluded from the regression. Because of the District's unique political and geographic characteristics, it is possible that its SAT results cannot be optimally fit using the same model used for the 50 states. If so, we would expect that including D.C. data in the regression would both lower the r-squared value and diminish the ability of the model to predict NAEP scores. This turns out to be the case. Hence, D,C, is omitted from the regressions.