Assessing School Effectiveness

Stephen Klein, Santa Monica, CA
David Freedman, Berkeley,CA
Richard Shavelson, Stanford, CA
Roger Bolus, Encinitas, CA

The Collegiate Learning Assessment (CLA) program measures value added in colleges and universities, by testing the ability of freshmen and seniors to think logically and write clearly. The program is popular enough that it has attracted critics. In this paper, we outline the methods used by the CLA to determine value added. We summarize the criticisms, which revolve around the question of which students take the CLA tests. Typically, samples are not random, so that selection bias is a concern, as is confounding. We respond by showing that criticisms of CLA procedures are not supported by the data.

Key words: Collegiate Learning Assessment, CLA, SAT, value added, regression, selection bias

Introduction

The Collegiate Learning Assessment (CLA) measures student skills in problem solving, analytic reasoning, critical thinking, and writing. The program focuses on these skills because they are applicable to a wide range of academic majors and are valued by employers. The program's goals are to provide colleges and universities with information about (1) how much improvement their students have made between the freshmen and senior years, and (2) whether that improvement is more or less than would be expected given the progress made by students at other schools. This information is intended to supplement rather than replace the way schools assess learning outcomes. Because CLA results are intended for internal use, the program does not publish individual or school level data.

One objective, then, is to measure "value added" by educational programs in colleges and universities—namely, the contribution each school makes to student learning in the areas tested. The CLA measures value added by comparing freshmen and seniors on tests that assess the skills described above. The CLA has attracted a great deal of attention,[1] and concerns have been expressed about its method for estimating value added. Here, we address those concerns.

We begin by describing how the CLA computes value added. Next, we summarize the criticisms, which revolve around the question of which students take the CLA tests. Typically, samples are not random, so that selection bias is a concern, as is confounding. We respond to the critics by showing that (1) the students who take the CLA tests are very similar to their classmates on SAT scores and other background characteristics; (2) participating seniors are very similar to participating freshman; (3) once there is control for SAT scores, performance on CLA tasks is not related to the content area of the task, student academic major, student demographic characteristics, or school characteristics like size. Thus, criticisms of CLA procedures are not supported by the data.

We then outline the research being conducted by the CLA, including the development of a new way to compute value added. This new method continues to use many of the key procedures employed by the old method, like cross-sectional comparison groups and covariate adjustments. The paper ends with a summary.

Current Procedures

Colleges and universities invite a sample of their freshmen and seniors to participate in the CLA program. Students who participate are often provided material incentives such as extra course credit, a gift certificate, or money. Alternatively, some schools embed the CLA in freshman writing or senior capstone courses. All the CLA tests are administered over the Internet and require open-ended responses. There are no multiple choice questions. The program uses six performance tasks and two types of essay questions ("make an argument" and "break an argument").

Because of time constraints, a student does either one performance task or two essay questions—one of each type. Although each student takes only a small portion of the test battery, all the questions are administered at each school. Questions are assigned randomly to students within a school, subject to balance conditions.[2] This procedure is called "matrix sampling."

Scores on the different CLA tests are converted to a common scale so that they can be combined across students to compute school means. The program informs schools about the difference in mean CLA scores between freshmen and seniors, and whether that difference is larger than, about the same as, or smaller than differences that are typically observed. CLA scores may be affected by differences in student ability prior to matriculation. Consequently, scores are adjusted for student ability.

The computation of a school's value added score utilizes the CLA scores and the test-takers' SAT scores or ACT scores if SAT scores are not available. ACT scores are converted to the SAT numerical scale using a table developed by the College Entrance Examination Board and adopted by virtually all college admissions officers (Dorans 1999). "Value added" is obtained as follows:

1) Regress mean freshman CLA scores on mean freshman SAT scores, with the school as the unit of analysis.

2) Use the equation from Step 1 to compute each school's residual, namely, its actual mean CLA score minus the "expected" value from the regression. (The expected value is usually referred to as the "predicted" or "fitted" value from the regression line.)

3) Repeat Steps 1 and 2 for the seniors.

4) A school's value added score is the difference between its residual for seniors and its residual for freshmen.

Value added scores are reported in five bands: well above, above, near, below, and well below expected. The percentage of schools in each band is preset at 10%, 20%, 40%, 20%, and 10%, respectively.

The freshman regression equation explains 79% of the variance in mean CLA scores across schools, whereas the senior equation explains 76%. The equations have nearly identical slopes. However, the senior equation has a larger intercept—because seniors generally earn higher CLA scores than freshmen. Klein et al. (2007) gives the rationale for the value added approach, as well as additional details on the CLA testing program. Also see Shavelson (2008).

Empirical results

We identified 93 schools that participated in the National Center for Education Statistics Integrated Postsecondary Data Systems (IPEDS) and had—

1) at least 25 freshmen taking the CLA in the fall of 2006 and at least 25 seniors taking it in the spring of 2007, each of these students having an SAT or ACT score, as well as complete data on age, race/ethnicity, gender;

2) complete data for the IPEDS school-level variables that will be used (including freshman median SAT, breakdown of undergraduates by race/ethnicity and gender, retention rate).

Data on CLA participant characteristics were obtained from a questionnaire completed by students before taking the test; SAT or ACT scores were obtained from registrars' offices. The empirical results reported below are based on these CLA and IPEDS data, unless otherwise noted. (Seventeen schools in the CLA program, enrolling 8% of the students, were excluded because data were missing on some variables.) To begin with, the left hand panel of Figure 1 shows a scatter diagram of mean CLA scores against mean SAT scores for freshmen; each data point represents one school. The right hand panel is for seniors. The senior data has the same shape as the freshman data, but the cloud of points is shifted upwards and a little to the right.

Figure 1.
Left hand panel: mean CLA score for freshmen plotted against mean SAT score of freshmen who took the test.
Right hand panel: mean CLA score for seniors plotted against mean SAT score of seniors who took the test.
CLA data: 93 schools.

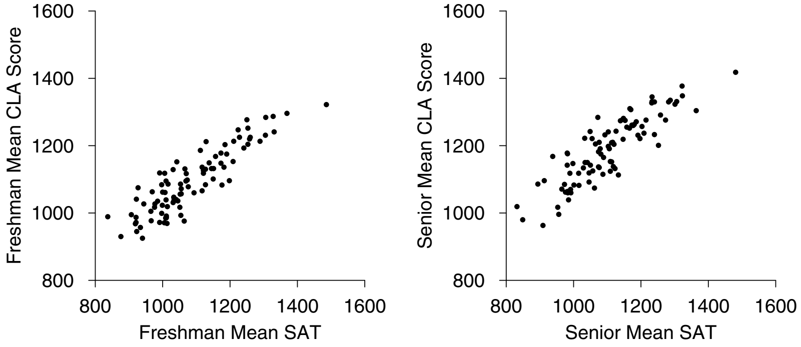


Figure 2.
Histogram for CLA value added scores: 93 schools.

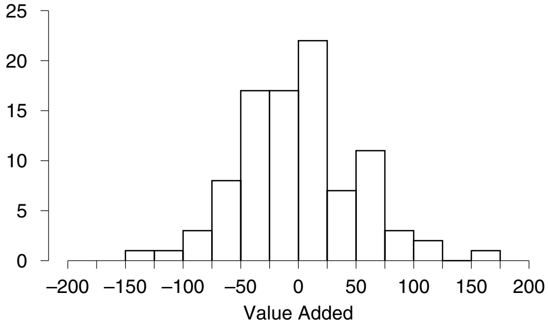
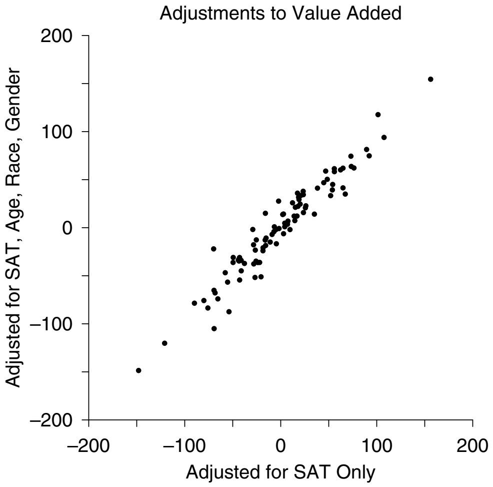


Figure 2 shows a histogram for the value added scores: the scale is set up so that value added is zero for a typical school.

Criticisms

This section describes the main criticisms of the CLA method for computing value added and our response to those concerns.[3]

Figure 3.
Value added adjusted for SAT, age, race, and gender
versus value added adjusted for SAT only.
CLA data, 93 schools.



Differences Between Schools. Students who participate in the CLA program at one school may, on the average, be stronger academically than those who participate at other schools. The CLA program uses the test-takers' SAT scores to control for such differences. However, students attending different schools may differ in ways that are related to CLA scores but not to SAT scores. This is a concern for some critics, because the SAT may not control for all the preexisting factors that cause differences in CLA scores across institutions.

To investigate that concern, we used the CLA data to construct a regression equation predicting a school's mean CLA score on the basis of (1) its mean SAT score and (2) its mean SAT score plus mean age, percent minority, and percent female. The unit of analysis was the school. Equations were constructed separately for freshmen and for seniors, resulting in two value added scores for each school—one adjusted for SAT, the other adjusted for SAT, age, race, and gender. The correlation across schools between the two value added scores was 0.96: see Figure 3. In short, after adjusting for SAT, other potential confounders do not matter.

We used the same strategy to investigate whether school characteristics mattered. Specifically, we replaced age, percent minority, and percent female as predictors with the following school-level IPEDS variables: number

Table 1.
Freshman mean SAT Score, Percent Minority, and Percent Female Among Those Tested in the CLA Program and Among All Students (IPEDS): 93 Schools.

	CLA	IPEDS	Difference In Means	Correlation between CLA and IPEDS
Freshman Mean SAT score	1084	1074	10	0.95
Percent minority	15.8	15.7	0.1	0.97
Percent female	62.6	57.4	5.2	0.66

Notes: Following the IPEDS definition, minority students are African Americans, Hispanics, or Native Americans. CLA percent minority and percent female cover the freshmen and seniors who took the test, whereas IPEDS data are for all undergraduates. IPEDS does not have usable data on age. IPEDS reports median SAT scores for freshmen only: the 1074 in the table is the mean of the 93 medians. CLA typically uses mean scores. The mean of the 93 CLA freshman mean SAT scores is 1084; the mean of the medians is 1079.

of full time equivalent students at the school, percent receiving Pell grants,[4] percent minority, an index of the school's selectivity,[5] whether the school is public or private, whether it grants doctoral degrees, whether it grants masters degrees. The correlation between the value added scores from the equations that did and did not include the IPEDS variables was 0.96. Thus, including the IPEDS variables in the equations had virtually no effect on a school's value added score once there was control on SAT scores. This is consistent with findings reported for students who took the CLA during the 2005–2006 school year (Klein et al. 2007). In short, differences between schools are unlikely to introduce bias.

CLA Participants Are Like Non-Participants. The CLA program encourages schools to test random samples of at least 100 freshmen and 100 seniors. However, most schools use convenience samples. Even with random samples, many students who are invited to participate will not do so, and those who actually participate may not be like a random sample from those who are invited to participate.[6] For such reasons, critics are concerned about selection bias.

We investigated this issue by comparing the SAT scores of the freshmen in the CLA program at a school with the SAT scores of all freshmen. We also compared the percentage of minority students among those tested in the CLA program with the percentage of such undergraduates at the school; likewise for females.

Table 1 shows that those participating in the CLA program are a lot like their classmates. For example, there was only a 10 point difference in the

Table 2.
Characteristics of Participating Freshmen and Seniors:
CLA Data, 93 Schools

	Freshmen	Seniors	Difference	Standard Deviation
Mean CLA score	1094	1191	96.9	96.6
Mean SAT score	1084	1104	20.0	121.4
Mean age	18.2	22.2	4.0	0.7
Percent minority	17.0	14.1	2.9	19.0
Percent female	62.5	63.5	1.0	11.9

Notes: Freshmen and seniors had standard deviations (SD of means across schools) that were nearly equal for each variable; the means of the two SDs are tabled. The SDs of percent minority and percent female in IPEDS were 18.1 and 11.6, respectively; the SD of freshman SAT was 123.8. There were 9057 freshmen and 6926 seniors in the CLA data. With the individual student as the unit of analysis rather than the school, the SDs would be considerably larger.

mean SAT scores, compared to a 97 point standard deviation (Table 2). With the school as the unit of analysis, there was a 0.95 correlation between the mean SAT score of the CLA freshmen participants and the mean SAT score of all freshmen at the school. The only difference of note is that women were somewhat more likely than men to participate. However, as reported above, if SAT is used to predict CLA scores, adding gender to the equation makes little difference.

Participating Seniors Are Like Participating Freshmen, Except for the Differences in Ages and CLA Scores. Some critics say it is inappropriate to compare seniors to freshmen because seniors who participate in the CLA program may be systematically different from the freshmen. To investigate that concern, we compared the freshmen and seniors on mean CLA score, mean SAT score, mean age, percent minority students, and percent females. Results are shown in Table 2.

The 97 point difference in mean CLA scores between freshmen and seniors corresponds to a whole standard deviation, whereas the 20 point difference between freshmen and seniors on the SAT corresponds to only 0.16 standard deviations. Thus, participating seniors look a lot like the participating freshmen, except that on the average they are 4 years older and have a much higher mean CLA score. Table 2 suggests that differential selection bias (between seniors and freshmen) is small. Controlling for SAT will make this bias even smaller. Results for the 2007–2008 academic year are virtually the same.

Differential Retention. Differences in retention rates are a potential source of differential selection bias between freshmen and seniors, and between schools. Academically stronger students are more likely to stay in college than their weaker classmates: seniors have higher SAT scores than freshmen (Table 2). As noted above, the difference is only 0.16 standard deviations, which cannot account for much of the one standard deviation difference between senior and freshman CLA scores, nor can it create much bias. Controlling for SAT scores reduces bias even further.

Interestingly, schools with stronger students tend to have higher retention rates than schools with weaker students: there is a correlation of 0.75 between a school's retention rate and its mean SAT score. Furthermore, the difference between a school's senior and freshman mean SAT scores is smaller at the schools with relatively high retention rates than it is at schools with relatively low retention rates.

Differential retention across schools tends to inflate estimates of value added for schools with low SAT scores, because it is the weaker students at these schools who drop out, while the stronger ones remain. That will inflate the mean CLA score of seniors at those schools. By contrast, differential retention does not seem to affect value added scores at schools with relatively high SAT scores.[7]

Intentional Selection. Some critics have argued that schools may try to stack the deck, for example, by choosing their best students to take the CLA tests. Table 2 shows this did not happen. Nor would it work. To stack the deck, a school would have to find freshmen who under-perform on the CLA (relative to their SAT scores) and seniors who over-perform—a tall order at best.

Longitudinal vs. Cross-Sectional Designs. Some critics have suggested replacing the CLA cross-sectional design with a longitudinal one. As we have indicated, biases in the cross-sectional design are likely to be small. Longitudinal designs do have advantages, but they also have drawbacks. For example, results may be stale by the time they can be reported. Panel bias is another drawback: participants lose interest and motivation.[8] Differential dropout is a third problem. The high cost of longitudinal studies should also be mentioned. On balance, the cross-sectional design seems preferable.

Matrix Sampling of Tasks. Some critics have objected to the fact that a student responds only to a sample of the CLA questions. However, the sample is balanced within schools. Thus, sampling is unlikely to introduce bias. Sampling does introduce random error. The amount of error will decrease as more students participate.

Interaction Between Task Content Area and Academic Major. Some of the CLA tasks (particularly the 90-minute performance tasks) may fit some academic areas better than others. Thus, how well a student performs on a

Table 3.
Prediction of CLA Performance Task Scores
Using Combinations of SAT score, Performance Task Area,
Student Academic Major Area, and Interactions.

Variables in the model	R^2
SAT	0.322
SAT + task	0.324
SAT + major	0.329
SAT + task + major	0.331
SAT + task + major + interaction of task and major	0.333

task may depend on the student's academic major. According to critics, that could be a problem.

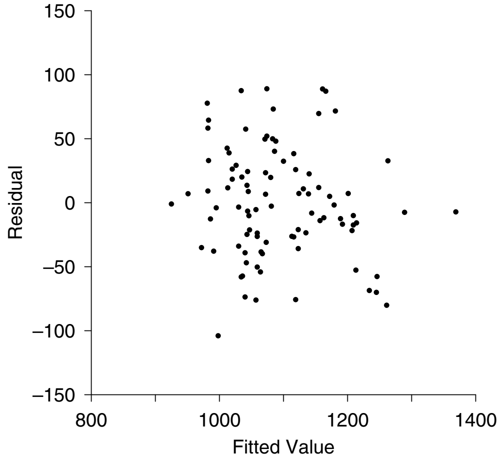
Shavelson (2008) investigated this concern using seniors who took a CLA performance task during spring 2007. He assigned each performance task to one of three content areas (science, social science, or humanities). Students self-identified the area of their major as science and engineering, social science, humanities, other. Finally, Shavelson constructed five student-level regression equations using combinations of SAT scores and dummies for task area and major area to predict CLA scores. Once SAT is in the model, other variables have almost no effect on predictive accuracy or value added (Table 3).

Maturation. Seniors with a given SAT score generally earn higher CLA scores than freshmen with the same SAT score. According to the critics, part of this difference is due to maturation rather than the school's educational program. In short, the CLA value added scores may overstate the benefits of the educational experience.

We agree that the difference between freshman and senior CLA scores may be affected by maturation. However, maturation would not affect the value added scores unless there are differences between schools in maturation rates. If students mature faster (in ways that affect their CLA scores) at some schools than they do at others, that difference should be credited to a school's value added score—which is what the CLA does.

Linearity and homoscedasticity. Questions have been raised about linearity and homoscdasticity. By way of example, consider the regression of freshman mean CLA scores (Y) on their mean SAT scores (X); the unit of analysis is the school. Data are shown in Figure 1. Let a be the intercept and b the slope of the fitted regression line, so the "fitted value"is $a + bX$ and the "residual" is $Y - a - bX$. Figure 4 plots residuals against fitted values. The data in Figure 1 seem quite linear; this is confirmed by the absence of

Figure 4.
Residuals vs fitted values,
from the regression of freshman mean CLA scores
on their mean SAT scores: CLA data, 93 schools.



any pattern in Figure 4. But there is a hint of heteroscedasticity—more variance at the left in Figure 4, less at the right.

To investigate this apparent heteroscedasticity, we computed an F statistic by ordering the fitted values from smallest to largest. The numerator of the statistic was the sum of squares of the residuals corresponding to the 18 smallest fitted values; the denominator was the sum of squares of the residuals corresponding to the 18 largest fitted values. The figure suggests that F should exceed 1. Indeed, $F = 1.19$.

Is this significant? To answer the question, we made a permutation test, randomly permuting residuals against fitted values to construct new Y's, computing F from the new data, and seeing what fraction of the new F's exceeded the F in the real data. This gave a one-sided P-value of 36%. We also used the parametric bootstrap, and got a one-sided P-value of 38%. The heteroscedasticity in Figure 4 is more apparent than real.[9]

Planned Research

Method for Computing Value Added. As explained above, the CLA program currently uses a linear regression model where the school is the unit of analysis and the school's mean SAT score is the sole explanatory variable. The CLA is examining another way of computing value added that may have certain advantages (Klein and Freedman 2008). This new method uses a regression model where the student is the unit of analysis and the predictors are the student's SAT score and a dummy variable for each school.[10] Additional characteristics also may be used as predictors in a sensitivity analysis.

There are separate equations for freshmen and seniors. The coefficient of a school's dummy variable is its "effect." These may be centered at an overall average. Value added by a school is computed as the difference between the effect on seniors and on freshmen. The new method has two advantages: (1) it gives a standard error for each school's value added score, and (2) it gives a residual score for each student that could be used in future research studies.

Transfer Students. Several schools in the CLA program have large numbers of both continuing and transfer students. Students who transfer typically do so between their sophomore and junior years. A school probably has a greater impact on those who took all their courses at the school from which they are graduating than those who took many courses at other schools. Transfers are therefore likely to dilute the school effect, biasing value added scores toward the overall average. The CLA program plans to conduct analyses at some of the schools with large numbers of transfers to assess the impact on value added.

Longitudinal vs. Cross-Sectional Designs. In 2005, the CLA program started a longitudinal study with 50 colleges across the country. This study will provide an opportunity to compare the longitudinal results at the roughly 35 schools that remain in this sample with the results from a cross-sectional design at those same institutions. The study will allow an assessment of the ways in which the two types of designs yield convergent or divergent estimates of value added.

Case Studies. The CLA program has begun on-site visits and interviews of students, faculty, and administrators at schools where the freshmen and seniors repeatedly score very differently from their expected levels; and at schools with unusually high or unusually low value added scores. These visits are intended to generate hypotheses about the sources of these differences. Possible explanations include student motivation and the nature of their high school academic program. This research may identify additional confounders that need to be considered. Or, it may help to resolve lingering concerns about selection bias.

Summary

1) Students who participate in the CLA program are very similar to their classmates on the dimensions we examined.

2) Except for being four years older (and scoring much higher on the CLA) participating seniors are a lot like participating freshmen.

3) The differences in background characteristics between participants and non-participants (like the differences between freshmen and seniors) are too small to matter—given the large difference between freshmen and seniors in CLA scores.

4) Adding different kinds of variables to a regression equation that includes SAT scores does not improve accuracy in predicting CLA scores, which reinforces points 1–3.

5) Although various kinds of selection bias are possible, the available data indicate that such biases are too small to matter (points 1–4).

6) Random selection of participants is not feasible—nor do we think it necessary, in view of point 5.

7) Longitudinal designs do not appear to be any better than the current cross-sectional design.

8) Like the concerns about selection bias, other concerns that have been voiced by critics (such as maturation, matrix sampling of measures, interactions, and stacking the deck) are unfounded.

Problems created by confounding and selection bias affect many kinds of research programs. The CLA is fortunate in being able to address these problems. There are two reasons. (1) There is a powerful control variable, namely, the SAT scores of test-takers. (2) There is enough data to assess hypotheses about confounding and selection bias that have been offered by the critics.

References

Dorans, N. J. (1999). Correspondences between ACT and SAT I Scores, College Board Research Report 99-1, College Entrance Examination Board, New York.

Freedman, D. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Freedman, D. and D. Lane (1983). A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics 1: 292–98.

Freedman, D., R. Pisani, and R. Purves (2007). *Statistics*. 4th ed., W.W. Norton, New York.

Klein, S. and D. Freedman (2008). Proposed method for measuring school effects and value added. Working Paper #1. Council for Aid to Education, New York.

Klein, S., R. Shavelson, R. Benjamin, and R. Bolus (2007). The Collegiate Learning Assessment: Facts and fantasies. Evaluation Review 31: 415–439.

Pascarella, E. T., T. Cruce, P. D. Umbach, et al. (2006). Institutional selectivity and good practices in undergraduate education: How strong is the link? The Journal of Higher Education 77: 251–85.

Shavelson, R. J. (2008). *The Quest to Assess Learning and Hold Higher Education Accountable*. Stanford University Press. To appear.

The authors

Stephen Klein is Director of Research and Development for the CLA program. email: steve@gansk.com

David Freedman is Professor of Statistics at the University of California, Berkeley. email: freedman@stat.berkeley.edu

Richard Shavelson is the Margaret Jacks Professor of Education at Stanford University. He is a consultant to the CLA. email: richs@stanford.edu

Roger Bolus directs statistical programming for the Council for Aid to Education, which sponsors the CLA. email: rbolus@netzero.net

Notes

1. In 2004–2005, 61 colleges and universities participated in the CLA program. In 2008–2009, over 300 are expected to participate.

2. The CLA test battery consists of six 90-minute performance test tasks, four 45-minute make-an-argument essay questions, and four 30-minute break-an-argument essay questions. Students are assigned either (1) to one of the six performance tasks or (2) to one make-an-argument and one break-an-argument essay question. Half the students are assigned to performance tasks, and half to essays. All students complete a background questionnaire. Additional information is available at the CLA website

http://www.cae.org/content/pro_collegiate.htm

There were minor deviations from protocol. For example, in some schools, students took both the performance and essay portions of the test.

3. Other criticisms of the CLA have been addressed by Klein et al. (2007).

4. Pell grants are need-based grants to low-income undergraduates.

5. From Barron's Magazine; provided by The Education Trust. See Pascarella et al. (2006).

6. Non-response is troublesome in many sample surveys (Freedman et al. 2007: 336, A20–21).

7. In our sample of 93 schools, there was a -0.27 correlation between a school's retention rate and the difference between senior and freshman mean SAT. Among the 20% of the schools with the highest retention rates, the difference between senior and freshman mean SAT scores was 2 points; among the lowest 20%, the difference was 38 points.

8. Freedman et al. (2007: 398) discusses panel bias in the Current Population Survey.

9. The cutpoint of 18 was chosen in conformity with note 7, as 20% of the number of schools. Changing the cutpoint to 10% or 5% makes little difference to the outcome. Results are similar for seniors, although significance is achieved when the cutpoint is 10%. The permutation test is explained in Freedman and Lane (1983); on the bootstrap, see for instance chapter 7 in Freedman (2005). The appearance of heteroscedasticity in Figure 4 may be due in part to a long right hand tail in the distribution of fitted values: there are fewer data points at the far right of the figure than the far left, so the vertical range is smaller.

10. This is a straightforward analysis of covariance model, with schools as treatments and SAT as the covariate. There are unequal numbers of subjects at each level of treatment. Such models are also called "fixed effects models." Value added may be computed separately for different components of the CLA test battery.