HOW TO ASSESS EXPECTED VALUE ADDED: THE CLA METHOD

Phillip Garcia

The goal for administering the Collegiate Learning Assessment (CLA) test is to estimate how well graduating seniors from a specific college or university perform with regard to critical thinking, analytic reasoning, problem solving, and written communication. The assumption is that students completing their undergraduate programs will perform better than an incoming group of undergraduates. Thus the primary focus of each CLA participant is the contribution of its institution to student improvement, or value added. Detecting whether program completers improved over time is accomplished by comparing CLA scores between separate samples of freshmen and seniors. A predictor variable, SAT (or ACT) score, is used to remove differences in Actual CLA test scores associated prior academic abilities. The benchmarks for each campus comparison are statistical parameters derived from a national database of participating schools. During the 2005-06 administration, there were 113 campus participants.

The current research design practice for the CLA project is not conventional; it compares sample observations from two cross-sectional groups rather than tracking one group's performance over time. The CLA authors (see Klein, Shavelson, Benjamin, and Bolus, 2007) provide no statistical proof for the derivation of their unorthodox method to detect value added, but they thoroughly document the parameters necessary to replicate their results. For example, all the regression statistics that underpin the campus-specific analyses are listed in Appendix C of a report entitled *Institutional Reports* (see http://www.cae.org). This is also the place where the CLA authors invite the reader to predict CLA scores for other alternative SAT scores.

For anyone that accepts the authors' invitation and then proceeds to systematically investigate the range of outcomes associated with the published regression parameters, it soon becomes clear that there is a slight defect in the approach used to amend CLA scores for prior differences in academic ability. On the surface things look good. The CLA authors begin by correctly using two separate regression equations to statistically control for the effects of prior differences; but what they fail to do is use a common parameter to simulate statistical equality between freshmen and seniors on their predictor variable. What follows are a detailed explanation of their error and the identification of more appropriate statistical adjustments. There is also a brief examination of the CLA research design and the sample designed employed by institutions using the CLA test.

UNEQUAL SLOPES

Let's start by specifying the null hypothesis for CLA performance level in traditional terms, that is, there is no difference between the two groups:

$$H_0\text{: } Actual\ Valued\ Added = Expected\ Value\ Added$$

An expanded version of this equality can be expressed as follows:

$$H_0\text{: } \{Actual(Y_{seniors}) - Actual(Y_{freshman})\} = \{Expected(Y_{seniors}) - Expected(Y_{freshman})\}$$

Here, the actual Y values signify the mean CLA scores for a cross-sectional sample of seniors and freshmen for a campus and the expected Y values signify the corresponding linear estimates associated with each of the actual mean CLA scores.

The two alternative hypotheses are expressed as follows:

$$H_1: \{Actual(Y_{seniors}) - Actual(Y_{freshman})\} > \{Expected(Y_{seniors}) - Expected(Y_{freshman})\}$$

$$H_2: \{Actual(Y_{seniors}) - Actual(Y_{freshman})\} < \{Expected(Y_{seniors}) - Expected(Y_{freshman})\}$$

The expression for H_1 indicates that the Actual Value Added is above expected performance level, and the expression for H_2 indicates that the Actual Value Added is below expected performance level.

To generate the campus-specific estimates of Expected Value Added for seniors, the CLA authors statistically rely on the parameters from two separate regression analysis. First they regress CLA scores (*Y*) on SAT scores (*X*) for freshmen observations, and then they do the same for senior observations. Next they generate two expected CLA scores by inserting the Actual SAT mean score from each group into its corresponding regression equation. The formula they use to predict each campus estimate is as follows:

$$\begin{aligned} Expected\ Value\ Added &= Expected\ (Y_2) - Expected\ (Y_1) \\ &= Y_2' - Y_1' \\ &= (a_2 + b_2\overline{X}_2) - (a_1 + b_1\overline{X}_1) \end{aligned}$$

where the subscript values *1* and *2* signify regression parameters derived from the separate samples of seniors and freshmen, respectively, where *a* is the actual intercept and *b* is the actual slope of the regression line derived from observations at all campuses, and where $\overline{X}_i$ signifies the actual mean SAT score for a specific campus. This procedure is meant to remove differences in *Y* associated with differences in *X*.

The problem with the above procedure is that the two expected values defined above do not share a common adjustment for SAT scores (i.e., $\overline{X}_2 \neq \overline{X}_1$). Two ways to attain predicted values that cancel out the effects of differing SAT scores is to either set the mean for seniors to the level observed for freshmen, or set the mean for freshmen to the level observed for seniors. The next two equalities show how these two options are expressed in statistical terms.

$$Expected\ Value\ Added = (a_2 + b_2\overline{X}_1) - (a_1 + b_1\overline{X}_1)$$

or

$$Expected\ Value\ Added = (a_2 + b_2\overline{X}_2) - (a_1 + b_1\overline{X}_2)$$

To illustrate the validity of the two adjustments cited above, let's compare examples of their empirical results with hypothetical results generated by the CLA authors' method. The most telling comparative results come from the special case where the mean SAT scores and mean CLA scores for both seniors and freshmen at a specific school equal the grand mean values observed across all participating schools during the 2005-06 administration. The comparative results using such values are displayed in table 1.

TABLE 1. Alternative Estimates of Expected Value Added

Row	Statistic	Freshmen	Seniors	Value Added
1	*a*	394	448	
2	*b*	0.652	0.690	
3	Actual SAT	1074.0	1100.0	
4	Actual CLA	1094.0	1207.0	113.0
5	Expected CLA (CLA Method)	1094.2	1207.0	112.8
6	*Expected CLA (X=1074)*	*1094.2*	*1189.1*	*94.9*
7	*Expected CLA (X=1100)*	*1111.2*	*1207.0*	*95.8*

Note: Regression parameters are from figure 1 and the means are from table 6, *Institutional Report* (n.d.).

The results from the CLA method are displayed in row 5 and the alternative results are displayed in row 6 and 7. The most noteworthy finding from table 1 is the equality between actual CLA scores and expected CLA scores for freshmen using the CLA method (compare rows 4 and 5). This equivalence is not an accident. Whenever the actual mean for the independent variable (i.e., *X*) is inserted into the regression formula for generating expected values of the dependent variable (i.e., *Y*), the result must be the mean for the observed dependent variable. The general expression is:

$$\bar{Y} = a + b\bar{X}$$

What the equality between actual CLA scores and expected CLA scores in table 1 reveals is that employing the CLA authors' method for generating expected values for CLA scores does not result in the anticipated adjustment for the below average SAT scores exhibited by the freshmen observations. The absence of any statistical adjustment confirms the inappropriateness of the CLA authors' method. The expected CLA score for freshmen should be higher than its actual CLA score because the freshmen averaged 26 points lower than seniors on the SAT.

The last two rows the show the derivation of Expected Value Added when inserting the statistically equalizes the two groups by entering the exact value of *X* into each regression equation. These two adjusted scores indicate that the correct magnitude of the expected is about 96 points. So, for this hypothetical example, the CLA method inflated Expected Value Added by 17 points (i.e., 113-96). Therefore, results from the 2005-06 administration might have mislabeled performance level regarding value added whenever freshmen and seniors at a campus differed noticeable in their SAT scores.

EQUAL SLOPES

Now how should the effect of prior differences be removed if the two observed regression slopes (i.e., 0.652 and 0.690) were statistically homogeneous? The traditional way to adjust for prior differences between two groups that separately produce equal regression slopes is to use the analysis of covariance (ANCOVA) technique, which employs a single regression line (see Kerlinger and Pedhazur, Chapter 10, 1973). Applying ANCOVA to the estimation problem at hand yields the following expression:

$$\begin{aligned} \textit{Expected Value Added} &= \textit{Adjusted}\ (Y_2) - \textit{Adjusted}\ (Y_1) \\ &= Y_{2(adj)} - Y_{1(adj)} \\ &= \{(\bar{Y}_2) - b_c(\bar{X}_2 - \bar{X}_c)\} - \{(\bar{Y}_1) - b_c(\bar{X}_1 - \bar{X}_c)\} \end{aligned}$$

where Y_i signifies the actual mean CLA score for either seniors or freshmen, X_i signifies the actual mean SAT score for both groups, X_c signifies the grand mean for the covariate (i.e., the mean SAT score derived from the pooled observations for both freshmen and seniors), and b_c signifies the common slope derived from the pooled observations of both freshmen and seniors.

TABLE 2. Expected Estimate of Value Added for Equal SAT Scores

Line	Statistic	Freshmen	Seniors	Value Added
1	*a*	421	421	
2	*b*	0.671	0.671	
3	Actual SAT	1074.0	1100.0	
4	Actual CLA	1094.0	1207.0	113.0
5	Grand Mean for SAT	1087.0	1087.0	
6	Adjusted CLA (ANCOVA)	1102.7	1198.3	95.6

The adjusted means scores for the two groups derived from the ANCOVA technique are displayed in the last row of table 2. Here, the freshmen mean is higher than actual because that group scored below the average of the pooled SAT scores; and the senior mean is lower than actual because that group scored above the average of the pooled SAT scores. The difference in expected CLA scores indicates the valued added when the two groups are statistically equal on the dependent variable. The difference in expected values is just less than 96 points, or essentially the same difference as generated by contrasting the two alternative estimates listed in rows 6 and 7 in table 1. The near identical differences confirm the appropriateness of the ANCOVA adjustment when slopes are essentially equal. And unlike the regression results, ANCOVA yields one set of unique adjustments

CROSS-SECTIONAL VS. LONGITUDINAL COMPARISONS

A secondary research focus for CLA institutions is the assessment of the relative status of campus freshmen to national norms and a separate assessment of the relative status of campus seniors to national norms. Here null hypotheses for freshmen and seniors is

$$H_0\text{: } \textit{Actual}(Y_{seniors}) - \textit{Expected}\ (Y_{seniors}) = 0$$

$$H_0: Actual(Y_{freshman}) - Expected(Y_{freshman}) = 0$$

Here it is entirely legitimate to use just the parameters from the two separate regressions to test the two suppositions cited above, as the CLA authors do. So CLA institutions can learn whether their freshman class is on par with the abilities of other campuses and then learn whether their seniors are on par too. These results, of course, do not reflect the notion of value added, but they do add context. The CLA authors use the following display to describe what can be learned from CLA participation.

III. 2005–2006 Institutional Results for Your School

University College	Freshmen	Seniors	Value Added
Mean SAT Score	1252	1250	
Expected CLA Score	1210	1311	101
Actual CLA Score	1170	1383	213
Difference (actual minus expected) *	-40	72	112
Difference (actual minus expected) **	-0.80	1.60	2.40
Performance Level ***	At	Above	Well Above

** In scale score points. ** In standard errors. *** Well Above, Above, At, Below, or Well Below Expected*

Freshmen: Based on the average SAT score (1252) of freshmen sampled at your institution, we would expect their average CLA score to be 1210. Your freshmen scored 1170, which is *At Expected*.

Seniors: Based on the average SAT score (1250) of seniors sampled at your institution, we would expect their average CLA score to be 1311. Your seniors scored 1383, which is *Above Expected*.

Value Added: Based on the average SAT scores of freshmen and seniors sampled at your institution, we would expect a difference of 101 points on the CLA. This difference is our estimate of the expected value added at your school. The difference between how your seniors scored (1383) and freshmen scored (1170) was 213 points, which is *Well Above Expected*.

Distribution of schools by actual minus expected scores (in standard errors) and performance levels

Because the mean SAT scores are basically the same for the 2005-06 freshmen and seniors, the above estimate of value added (i.e., 112) is reasonable. So the results indicate that an "average" freshmen class was transformed into an "above average" senior class because the value-added experience at the institution was "well above average."

Another nagging thought is that seniors at many institutions may represent a special subset of all incoming freshmen: they are the survivors of the undergraduate curriculum. This probably is not the case for highly selective institutions that have graduation rates that exceed 80 percent; but it is a concern for less selective institutions that have much lower completion rates.

At this time, 45 CLA institutions are piloting a longitudinal assessment of a freshmen cohort. The use of longitudinal observations may still require the use of predictor variables or covariates to reduce the observed variability; but, certainly, the tracked observations will be more robust than the cross-sectional observations when it comes to the standard threats to validity, like selection bias, history, repeated testing, and regression to the mean (see Campbell and Stanley, 1966).

Because of attrition, samples sizes for the longitudinal observations will have to be larger than the minimum samples size recommended for the analysis of cross-sectional data. Of course, larger numbers longitudinal observations will not automatically take care of issues like selection bias or experimental mortality. For example, if incoming students are selected from convenient natural settings (e.g., orientation classes), will the selected students truly represent the various academic paths students follow at an institution? Will differential time-to-degree patterns affect test performance? Moreover, will it even be possible to capture all the sampled students that attain senior class status? How to successfully select and a random sample of student for voluntary testing and then tracking those students for retesting four years later is something that still needs to be documented.

SAMPLE SIZE

The published results from the 2005-06 Institutional Report suggest that most of the participating schools appear to have sampled between 100 and 200 freshmen and a similar number of seniors. So even if we assume that both the freshman and senior samples originated from populations that averaged 1100 on the SAT, they each would still have sizeable margins of error around their SAT estimates. If the standard deviation for the SAT were 130, the 2005-06 observation for seniors, the error margins would range from a low of $\pm$18 SAT points to a high of $\pm$24 SAT points. Of course, if the two cross-sectional groups originated from populations that did not share a common SAT mean score, then the probability of differing average SAT scores between freshmen and seniors would be much higher than the comparable probability associated with random error.

TABLE 3. 95% Confidence Interval When Average SAT Equals 1100 for Selected Degrees of Freedom

df	*t*-values for α=.05	STDERR	95% Confidence Interval	
			Minimum	Maximum
10	2.23	43	1004	1196
20	2.09	30	1037	1163
30	2.04	24	1051	1149
40	2.02	21	1058	1142
60	2.00	17	1066	1134
120	1.98	12	1076	1124
200	1.96	9	1082	1118
300	1.96	8	1084	1116
400	1.96	7	1086	1114
500	1.96	6	1088	1112
600	1.96	5	1090	1110
1200	1.96	4	1092	1108
2300	1.96	3	1094	1106

Note: Adapted from Downie and Heath, Appendix Table III (1965).

Like all sample-based parameters, campus-specific CLA scores would probably benefit from larger rather than smaller numbers of students. Certainly campuses should continue to sample at least 100 students for each comparison group; and they probably should be diligent to obtain equal numbers of freshmen and seniors. Naturally, this all-purpose recommendation applies to comparison groups derived from simple random sample procedures or stratified random sample procedures, that is, when every targeted student has an equal chance to be selected. It does not apply to samples derived from cluster designs. If CLA institutions are selecting freshmen and seniors from specific class sessions, then they are employing a cluster design. The main benefit of the cluster design is reduced cost for gathering data. The disadvantages are a possible bias in selection and larger sampling errors. One study (Chatman, 2007), for example, warns researchers that use aggregate measures like the CLA test are sensitive to the presence or absence of certain students. Data from the Undergraduate Survey administered to students at the University of California suggest that students in the Humanities would be more likely to score above average on the CLA and students in the physical sciences would be more likely to score below average. Regarding the loss of statistical precision, the rule-of-thumb formula for estimating the design effect of a cluster sample mean (Kalton, 1983) indicates that the standard error would be at least three times higher for the cluster sample than for a simple random sample, with the same number of observations. Looking back at table 3, the estimated standard error for the SAT mean based on 120 observations is 12 SAT points. If the same estimate were drawn from 6 class sessions that each contained 20 students, the standard error probably would be greater that 36 SAT points.

CONCLUSIONS

There is continuing pressure from multiple fronts for higher education institutions to regularly measure student learning with standardized instruments. It could be that the statistical error present in the analysis of 2005-06 CLA outcomes is one indicator of how CLA researchers now might be moving too fast to satisfy the external demands for assessment. The miscalculation in how to estimate Expected Value-Added was certainly not an obscure error. All anyone needed to do to detect and correct the problem was check any classic text on how to compare two separate regression slopes or how to execute an analysis of covariance. It may not be so important that the CLA researchers made the original error; mistakes happen. But it is important that the error went undetected by any of the 130 institutions that administered the CLA test.

Immediate use of cross-sectional approach to assessing value added is another indicator that CLA researchers maybe moving too fast. Its primary asset seems to be that it will generate CLA results in the shortest possible observation period; and its other asset seems to be the ease of the test administration. In contrast, the longitudinal approach will take much longer to yield terminal findings and it probably is the more difficult way to collect CLA test results. Moreover the longitudinal approach is not free from all the threats of validity that plague the cross-sectional approach. But, in the long run, the scientific literature suggests the longitudinal approach will produce the more robust findings on how much students learn as they move from college freshman to college senior status.

The last indicator of haste is the lack of information on how cross-sectional samples were drawn at the 130 institutions that gathered CLA data in 2005-06. It is not clear from the CLA documentation found on the Internet (http://www.cae.org) that CLA researchers provided guidelines for selecting students or that campuses adhered to any acceptable options. The details of sampling designs should be of interest to those that collect cross-sectional data. Such details, however, would be of less interest to those that focus on bivariate relationships derived from longitudinal data.

APPENDIX: Graphic Representations of Expected Added Value

The graph in figure A1 graphically illustrates the three estimates of Expected Value Added when the SAT-CLA regression slopes between freshmen and seniors are assumed to be different. Here, each estimate of Expected value Added is represented by a distance between the separate regression lines for freshmen (bottom) and seniors (top). The estimate derived by the CLA method is the diagonal dotted line that intersects the two regression lines at the expected *Y* value for freshmen (1094) and expected *Y* value for seniors (1207). This is the flawed estimate. The telling sign is that the dotted line does not vertically traverse the two regression lines. The other two estimates are the more reasonable ones. The estimate on the left reflects the assumption that both groups had equal SAT scores at the level observed for freshmen (i.e., 1074); the estimate intersects the freshmen line at *Y'*=1094 and the senior line at Y'=1189. The estimate on the right reflects the assumption that both groups had equal SAT scores at the level observed for seniors (i.e., 1100); the estimate intersects the freshmen line at *Y'*=1111 and the senior line at *Y'*=1207.

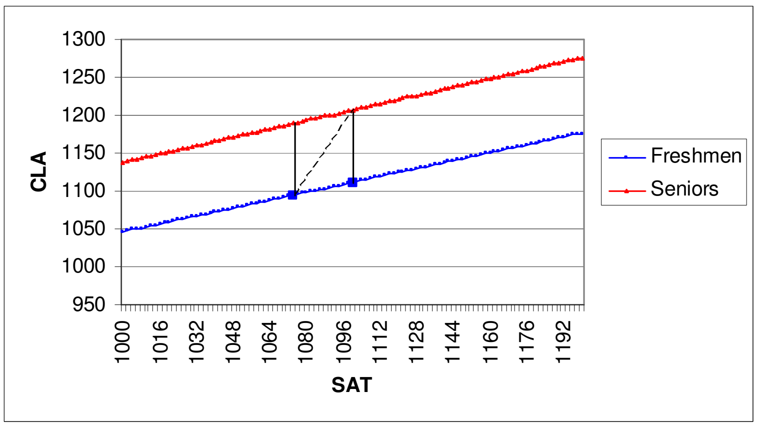


Fig. A1. Three Regression Estimates of the Expected Value Added

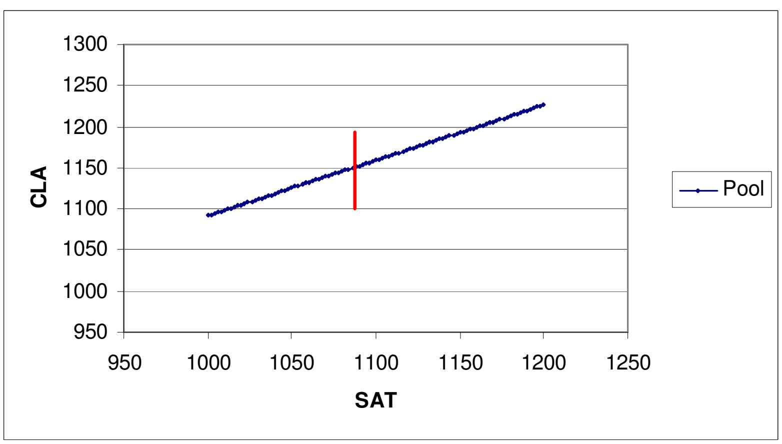


Fig. A2. ANCOVA Estimate of the Expected Value Added

The graph in figure A2 illustrates the Expected Added Value generated by the analysis of covariance method, which assumes that the regression slopes between freshmen and seniors are statistically equal. The plotted points represent the regression line generated the pooled observations of freshmen and seniors. The vertical line represents the Expected Value Added, where the mean of X is equal to the pooled estimate for all freshmen and seniors. The top endpoint of the vertical line reflects the $Y_{adj.}$value for seniors (i.e., 1198) and the bottom endpoint reflects the $Y_{adj.}$ value for freshmen (i.e., 1103).

REFERENCES

Campbell, D. T., and Stanley, J. C. (1966) *Experimental and Quasi-experimental Designs for Research*, Chicago, Illinois: Rand McNally.

Chatman, S. (2007) Institutional versus academic discipline measures of student experience: a matter of relative validity, *Research & Occasional Paper Series*, Berkeley, California: University of California, Berkeley.

Council for Aid to Education, (n.d.) *College Learning Assessment: Institutional Report*, retrieved June 16, 2007, from http://www.cae.org.

Downie, N.M. and Heath, R.W. (1965) *Basic Statistical Methods*, New York, NY: Harper and Row.

Kalton, G. (1983) *Introduction to Survey Sampling*, Newbury Park, California: sage Publication.

Kerlinger, F.N. and Pedhazur, E.J. (1973) *Multiple Regression in Behavioral Research*, New York, NY: Holt, Rinehart, and Winston, Inc.

Klein, S, Shavelson, R., Benjamin, R, and Bolus, R. (n.d.) *The Collegiate Learning Assessment: Facts and Fantasies*, retrieved June 16, 2007, from http://www.cae.org/content/pro_collegiate_reports_publications.htm.

Comment on
Philip Garcia's Critique of the CLA:
"How to Assess Expected Value Added: The CLA Method"

Richard J. Shavelson
Stanford University

Philip Garcia rightly raises a concern about the CLA's method for estimating value added. In essence, he argues that the most appropriate model for estimating value added is the standard analysis of covariance. This contrasts with the CLA's approach in which the difference between the value-added estimate for freshmen is subtracted from the value-added estimate for seniors. If, among other assumptions, freshmen and seniors have the same mean ACT/SAT scores in the population at each of the campuses assessed, his approach would be the appropriate one. However, the assumption of equivalence may be problematic. It turns out that Mr. Garcia's recommendation is a special case of the CLA approach, the approach that would be taken when the assumption of equivalence is tenable.

More specifically, Mr. Garcia's position is that the CLA should use a single regression equation for estimating value added, not the difference between two equations that the CLA currently uses. The expected value added by any particular campus, then, would be the difference between freshmen and senior mean scores adjusted for the common mean ACT/SAT scores. Among other assumptions, this analysis of covariance approach assumes that the freshmen and seniors are representatively sampled from the same population and any differences between mean ACT/SATs of the two groups arises only by chance. Hence an average the two groups' ACT/SAT scores would provide a better estimate of the common population mean than either one, and the single-equation analysis of covariance would make the appropriate adjustment (*all else equal*).

However, the underlying assumption of equivalence between freshmen and seniors on their ACT/SAT scores is quite likely untenable in most practical applications of the CLA (and the MAPP and the CAAP) for two reasons. First, student recruitment to take the CLA on each campus is not done randomly. Rather, volunteers are recruited and their recruitment is very difficult. Consequently, it is quite possible that the samples of freshmen and seniors taking the CLA are not representative of all freshmen and seniors on the campus. Second, due to churning of students between their freshmen and senior years—with some students dropping or stopping out and others entering in between freshman and senior years—we should not necessarily assume that ACT/SAT scores balance out on average and the freshmen and seniors have comparable mean scores. Indeed, as ACT/SAT predicts GPA in most colleges, we might reasonably expect, on balance, that the seniors would represent a more selective population than the freshmen.

If the assumption of equivalence of freshman and senior populations holds, the analysis of covariance would provide a better model for value added than the current one used by the CLA. However, there is good reason to suspect the assumption does not hold. Consequently the analysis of covariance does not make the proper adjustments for value added. Rather, the CLA's use of the difference in expected value added recognizes that the mean ACT/SAT scores probably differ in the population and so estimates value added for freshmen and seniors separately and then takes the difference.

REPLY TO SHAVELSON'S RESPONSE

The CLA method of using cross-sectional data to generate value added scores is so unorthodox that it does not fall into any of the categories normally associated with quasi-experimental or non-experimental research designs. So people that promote or critique the method cannot rely entirely on codified terms or symbols when they converse. In hindsight, I confess that I might have presented my case better if I had not categorized my reanalysis of CLA data as results from "two separate regression equations" and results from an "analysis of covariance." Both of these expressions denote very specific notions about who is being observed and the nature of their regression slopes. Unfortunately, some of the notions truly do not apply to the CLA method. So I take some responsibility for Dr. Shavelson's overall assessment that the gist of my paper was to suggest that an analysis of covariance was the better statistical technique for detecting value added scores from extant CLA data. That was not my intent. What I had hoped to do was show that the CLA method does not generate valid scores for expected value added scores and I offered alternative results from both an analysis of separate regressions and an analysis of covariance to make my point. Since that tack proved to be unsuccessful, I think it best now that I simply restate my critique, excluding any references to the use of separate or common regression equations. I stand by my original conclusion. But this time around, I want to make clear up front the real bottom line: as a rule, cross-sectional data from unrelated cohorts of freshmen and seniors cannot generate valid estimates for a complete analysis of value-added scores for the CLA test or any other test.

THE PROOF

I'll begin my discussion of valid estimates by restating how to detect value added scores when the appropriate data have been collected, longitudinal observations; and then discuss the statistical adjustments that might be made when longitudinal data are not available. Like before, my starting point is the null hypotheses for value added scores. The null hypothesis for longitudinal observations collected at two time points is

$$H_0 = \text{Actual (Posttest)} - \text{Actual (Pretest)} = \text{Expected (Posttest)} - \text{Expected (Pretest)} \qquad \textit{eq. 1}$$

That can be expressed symbolically as

$$H_0 = Y_2 - Y_1 = Y'_2 - Y'_1 \qquad \textit{eq. 2}$$

where Y' equals an expected score and the subscripts represent the before (1) and after (2) time points. If the parameter X equals an SAT/ACT score for an institution, then the expected added value derived from the pretest- and posttest can be expressed as

$$Y'_2 - Y'_1 = \{a_2 + b_2(X)\} - \{a_1 + b_1(X)\} \qquad \textit{eq. 3}$$

Now, by substitutions, the null hypothesis can be expressed as the equality between an observed difference and an expected difference based on two separate regression lines.

$$H_0 = Y_2 - Y_1 = \{a_2 + b_2(X)\} - \{a_1 + b_1(X)\} \qquad \textit{eq.4}$$

For me, eq. 4 represents the analytic standard for detecting value added.

The problem at hand has been that the CLA data collected to date do not accommodate usage of the analytic standard. Because the pre-test and posttest scores have been cross-sectional observations, the extant CLA data on seniors have not include observations for Y_1 and the data collected for freshmen do not include observations for Y_2. So the challenge the CLA team accepted was to find reliable estimates for either the two pretest scores, Y_1 and Y'_1 for the current crop of senior. I do not believe the CLA team always made the best choices.

Before comparing how the CLA team handles estimates for the missing pretest data for seniors versus how I would handle it, I want to make it assure the reader that the value added scores derived by the CLA method did emerge from a computational formula that mimics the same computational formula denoted by eq. 1. For the record, the CLA version of eq. 1 is sometimes represented as

$$H_0 = \text{Actual (Posttest)} - \text{Expected (Posttest)} = \text{Actual (Pretest)} - \text{Expected (Pretest)} \qquad \textit{eq. 5}$$

But this, by itself, is not a problem. In two steps the terms in eq. 5 can be rearranged to look exactly like eq. 1. So the CLA method approached the missing data problem the same way I did: it specified an analytical standard and then sought to enter surrogate measures for the two pieces of absent information. It did not seek to remedy the situation by injecting any new variables into the analysis.

First, here is how they chose to estimate Y'_1 at $time_1$ for seniors with cross-sectional data from freshmen observations taken in the same academic year. Using the subscript *s* to signify seniors and subscript *f* to signify freshmen, the equality is

$$Y'_1 = \{a_s + b_s(X_s)\} - \{a_f + b_f(X_f)\} \qquad \textit{eq. 6}$$

The alternative I put forward in my original paper was

$$Y'_1 = \{a_s + b_s(X_s)\} - \{a_f + b_f(X_s)\} \qquad \textit{eq. 7}$$

Both equations assume that the regression parameters for freshmen observed at $time_1$ are the best linear predictors for the unobserved regression parameters for seniors. This is not a leap of faith. We both believe that the general expression $a_f + b_f(X)$ has predictive validity. T he latest graphic posted by the CLA team shows how stable the *xy*-regression line is for freshmen across cohorts (see figure A in the appendix). The difference is that my equation (eq. 6) assumes that the best X estimator for generating an expected pretest CLA score for seniors should be the observed SAT/ACT score for seniors. My choice is based on the logic of the analytic standard. In eq. 4, the value of X cannot change over time, so the value of X must be the same for both the Y'_1 term and the Y'_2. It follows, then, that the value of X should be the same in both terms for any alternative equation one eventually chooses to employ.

Now, what surrogate measure did the CLA team select for the Y_1, the observed pretest for seniors? It was the observed CLA score for freshmen. Thus the CLA score for new undergraduates was assumed to be the same as the CLA scores for students that entered the university as new undergraduates at least 3.5 years earlier. I cannot think of any rationale for making that assumption for the range of schools involved in the CLA project.

To start, the test data collected by the CLA team do not support that assumption. If, for instance, the pretest scores for the two cross-sectional groups were essentially the same, then you would expect their SAT scores should be essentially the same. They were not. For the first CLA administration the mean for observed seniors at all schools was, on average, 44 points higher than the comparable mean observed for freshmen. In the second administration, the difference at all schools was, on average, 37 points, again, with freshmen exhibiting the lower score.

I can think of some institutions where the assumption that pretest scores for freshmen in a fall term might be a good surrogate of pretest scores for a contemporary group of seniors. They would be elite schools that enroll students from a lofty restricted range of SAT scores that also exhibited six-year graduations rates at 80 percent or above. Perhaps a small, but elite liberal arts school would be the perfect example. However, I cannot imagine the assumption holding up for most of the less selective schools (SAT < 1000 present in the CLA administrations, or the moderately selective schools (i.e., 1000 < SAT < 1150). At these institutions the standard deviation for SAT scores can be larger than those observed among the freshmen that took the CLA at either administration (i.e., around 145 points). For example, at the 23 campuses of the California State University, the standard deviation for SAT scores at less selective campuses has been around 165 points, and around 155 points for moderately selective campuses. Moreover, the graduation rates less selective and moderately selective schools have been between 40 and 60 percent. At that range, it would be hard to argue that seniors represent a random sample of incoming freshmen.

Now let's go back to the standard (eq. 4). If the data represented longitudinal data, of course, the following formula would be a valid expression for estimating the true value of Y_1

$$Y_1 = \{a_2 + b_2(X_2)\} - \{a_1 + b_1(X_1) - Y_2 \qquad \textit{eq. 9}$$

But when the data are cross-sectional, neither the CLA adjustment for Y'1 (eq. 6) nor my adjustment for Y'1 (eq. 7) yields a valid estimate of Y_1. Thus

$$Y_1 \neq \{a_s + b_{s1}(X_s)\} - \{a_f + b_f\,(X_f) - Y_2 \qquad \textit{eq. 10}$$

and

$$Y_1 \neq \{a_s + b_s(X_s)\} - \{a_f + b_{f1}(X_s) - Y_2 \qquad \textit{eq. 11}$$

Bottom line: if the cross-sectional data at hand do not yield a valid estimate Y_1, then the same data cannot yield a valid estimate of value added.

Valid Estimates

At best the cross-sectional CLA data collected so far can offer a partial glimpse of value-added. I'll illustrate what you can estimate using the statistics posted for the total sample in the 2006-07 CLA administration. They are displayed in the table 1 (see below). The first two rows list the linear parameters for the two observed regression lines; the third and fourth rows list the observed SAT scores and the observed CAL scores for all institutions.

Table 1. 2006-07 Sample Statistics for All Institutions

Statistic	(1) Freshmen	(2) Seniors
a	346	397
b	0.69	0.72
Mean SAT Score	1067	1104
Actual CLA Score	1057	1243

Source: http://www.cae.org/content/pdf/CLA_2006-2007_Sample_Institutional_Report.pdf, table 9, page 16; retrieved on November 19, 2007.

The values listed in table 2 (see below) are derived directly from the values listed in table 1. The first row in table 2 lists the two expected CLA scores generated by eq. 6 and their difference. The expected value added equals 84. The second row lists the observed CLA score for seniors, and the third row lists the difference between the observed CLA score for senior and the expected CLA score for seniors. The figure 51 at the bottom of column 3 indicates that the actual seniors CLA score exceeds the expected score. As I said, the results are limited, but they are all that the cross-sectional data can legitimately bear.

Table 2. Valid Estimates for Expected Value Added and
Actual versus Expected CLA for Seniors

	(1) Freshmen	(2) Senior	(3) Value Added
Expected CLA Score	1108	1192	84
Actual CLA Score		1243	
Actual versus Expected		51	

The last table I want to review lists hypothetical results using the standard CLA method. The numbers in table 3 (see below) were derived from the parameters listed above in table 1 and the format of table 3 mimics campus reports that CLA distributes to its each of its participants. My first comment is that the assessment of actual versus expected CLA scores are completely legitimate for the two cross-sectional groups. The -25 at the bottom of the freshmen column indicates that the freshmen scored below expectation and the score of 51 shows that the seniors score higher than expected. In contrast, the strikethrough values listed in column 3 represent illegitimate estimates of value added. The estimate of expected valued added, 110, is erroneously high because there was a large gap in SAT scores between the observed freshmen and observed seniors (1067 versus 1104). The estimate 186 is erroneous because there is no foundation for assuming that the cross-sectional freshmen CLA score (1057) is the best estimate for the unobserved pretest scores for seniors. And the difference between actual value added and expected value added, 76, is erroneous because it is based on two flawed estimates.

Table 3. A Sample of the Standard CLA Output using
2006-07 Sample Statistics for All Institutions

2006-07 Summary Data	(1) Freshmen	(2) Seniors	(3) Value Added
Mean SAT Score	1067	1104	
Expected CLA Score	1082	1192	~~110~~
Actual CLA Score	1057	1243	~~186~~
Actual versus Expected	-25	51	~~76~~

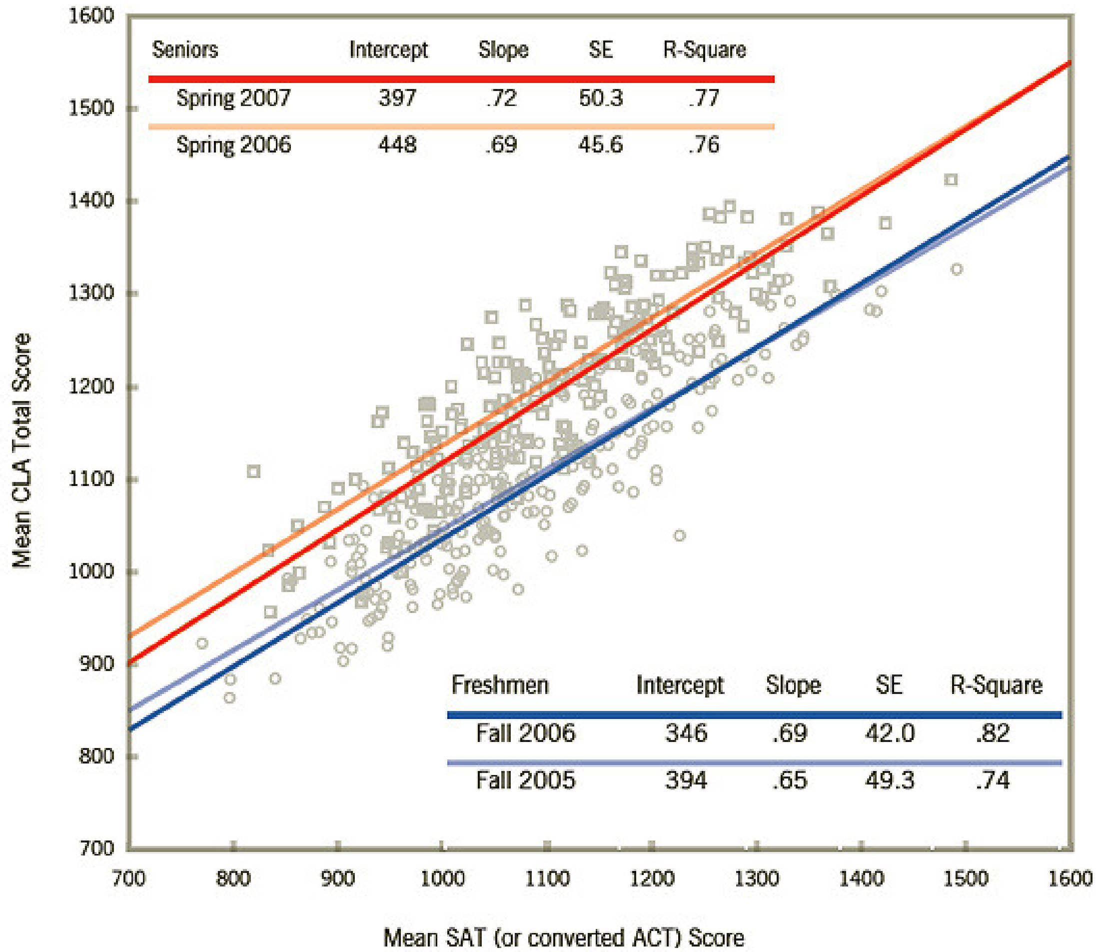
FINAL REMARKS

For the 2005-06 CLA administration, the two separate regression lines the CLA web site currently posts (see the lighter lines in the appendix) denoted a set of expected value added scores that ranged from 73 to 96 CLA points. When a school's SAT score was around 1073 (the average for the 2006-07 data), the expected value added was about 83 CLA points. When I looked at some of the examples that were posted for individual schools with more or less average SAT/ACT scores, the expected values that were listed exceeded 100 CLA points. That inconsistency caught my eye as a red flag. If the regressions had been based on longitudinal data, all the estimates for the expected value added scores should have landed between 73 and 96 CLA points. So I began my investigation on how expected value added scores should be determined. My first conclusion was that the CLA method generates inflated estimates of expected value added when the cross-sectional groups differed on their observed SAT scores. Naturally, the larger the difference observed between groups, the greater the distortion. Upon further investigation, I discovered other outcomes derived from the CLA method were problematic too.

The big ticket item for me was never whether to use analysis of variance or separate regression equations to detect expected value added with collected CLA data. That problem can be easily resolved by dividing observed slope differences by an appropriate standard error. If the ratio is greater than 1.96, use separate regression lines; if not, use analysis of covariance. Rather the big ticket item was always whether to use longitudinal data or cross-sectional data in the assessment of value added scores. I see no reason to ignore the conventional wisdom: longitudinal data fit the bill better than cross-sectional data. When longitudinal data are applied to the problem of detecting value added, then the one-group pretest-posttest pre-experimental research design describes the analytic approach one should use. The CLA team should resume their complete analysis of value added when the necessary longitudinal are available.

APPENDIX

Figure A: Relationship between CLA Performance and Incoming Academic Ability



Source: http://www.cae.org/content/pdf/CLA_2006-2007_Sample_Institutional_Report.pdf, page 25; retrieved on November 19, 2007.