Improving the Reliability and Interpretability of Value-Added Scores for Post-Secondary Institutional Assessment Programs

Jeffrey T. Steedle
Council for Aid to Education

Presented at the 2010 Annual Meeting of the American Educational Research Association
Denver, CO
May 3, 2010

Abstract

Tests of college learning are often administered to obtain value-added scores indicating whether score gains are below, near, or above typical performance for students of given entering academic ability. This study compares the qualities of value-added scores generated by the original Collegiate Learning Assessment value-added approach and a new approach that employs hierarchical linear modeling (HLM). Results indicate that value-added scores produced by the two approaches are correlated and would be identical if larger student samples were available. The new approach produces value-added scores that are slightly more reliable and display substantially greater consistency across years. Furthermore, the HLM-based approach provides school-specific indicators of value-added score precision, which improve interpretability. For these reasons, the new approach is recommended for future institutional assessment programs.

Improving the Reliability and Interpretability of Value-Added Scores for Post-Secondary Institutional Assessment Programs

Institutional value-added scores intend to capture whether the growth in performance on a standardized test of college learning between freshman and senior year is below, near, or above what is typically observed at schools testing students of similar entering academic ability. The estimation of value-added scores for post-secondary institutions is a relatively new enterprise, so few studies have evaluated or compared the statistical properties of alternative value-added estimation approaches. At present, there is one approach that predominates in higher education: the ordinary least squares (OLS) regression method first used by the Collegiate Learning Assessment (CLA) for its 2004-2005 test administration. This approach is also currently used by ETS and ACT to generate value-added scores for the Voluntary System of Accountability College Portrait (Keller & Hammang, 2008).

The original CLA approach involves computing the difference between senior and freshman residual scores based on regressions of mean CLA scores on mean SAT scores. Prior research demonstrated that value-added scores generated by this approach are reliable (Klein, Benjamin, Shavelson, & Bolus, 2007) and that students taking the CLA are generally representative of the larger student populations from which they are drawn (Klein, Freedman, Shavelson, & Bolus, 2008). Thus, the original approach should provide reasonable value-added score estimates.

That said, one could imagine a value-added estimation approach producing scores with even greater reliability (or equivalently less error). The original CLA approach is based on the average difference in performance between freshmen and seniors, but difference scores tend to be less reliable than the scores from which they are derived (Crocker & Algina, 1986). A value-

added estimation approach that does not depend on difference scores may provide better value-added score reliability.

Moreover, an increase in reliability would likely improve the year-to-year consistency of value-added scores, which some schools perceive as unrealistically low. Of course, value-added scores should not be identical across years (e.g., due to programmatic changes, major differences in sampling, or measurement error), but they should not change radically either. Substantial value-added score variability over time diminishes the ability to interpret differences across years, and this would create problems for an assessment program that intends to stimulate improvements in teaching and measure subsequent impacts on learning. For this reason, it is essential that a value-added estimation approach produce scores that do not vary over time in unrealistic or inaccurate ways.

Information about score precision serves as a reminder that value-added scores are estimates with inherent uncertainty. The original CLA value-added approach provides a single index of score precision that is used to characterize the uncertainty of all schools' value-added scores. An improved approach should provide school-specific indicators of precision because precision varies from one school to another depending, for example, on the sample size of students taking the CLA. Acknowledging this uncertainty would facilitate honest interpretations by providing a realistic sense of the variation in value-added scores that should be expected if different samples of students were tested (e.g., with a 95% confidence interval). Information about score precision could also be used to determine what constitutes credible and trustworthy differences in value-added scores. Furthermore, precision improves as sample size increases, so schools would be rewarded with greater precision and therefore greater interpretability of value-added scores by meeting or exceeding sample size targets.

This report presents analyses comparing the original CLA value-added estimation approach to a new hierarchical linear modeling (HLM) approach that could provide the improvements described above. Analyses were carried out on data from recent administrations of the CLA, a measure of college students' critical thinking and written communication skills as applied to authentic, open-ended problems (www.cae.org/cla). Results of these analyses will inform decisions about the value-added estimation approach used by major testing organizations for institutional assessment programs in higher education. The following sections provide background on these two approaches, a description of methods for comparing them, the results of comparative analyses, and a discussion of results.

The Original Approach

To avoid the expense of testing all students, schools typically administer the CLA to roughly 100 freshmen during the fall and 100 seniors during the following spring.[1] At nearly all schools, seniors outperform freshmen on average, but the average freshman-senior difference varies widely across schools. For a given school, the value-added score estimated by the original approach indicates the degree to which the observed average freshman-senior difference is below, near, or above expectations, where expectations reflect the freshman-senior difference one would expect given the average entering academic ability of test takers. This process, which is based on OLS regressions of mean CLA on mean SAT (or converted ACT) scores, is depicted in Figure 1 for a fictional school called University College. Schools at which the freshman-senior differences exceed expectations are said to have high "value added" because students attending

[1] Note that this sort of data collection (testing different groups of freshmen and seniors during the same academic year) provides a "cross-sectional" value-added estimate. Although it is an option for schools participating in the CLA, very few schools are willing to commit the funding and resources required to carry out longitudinal data collection (or wait 4 years to get results).

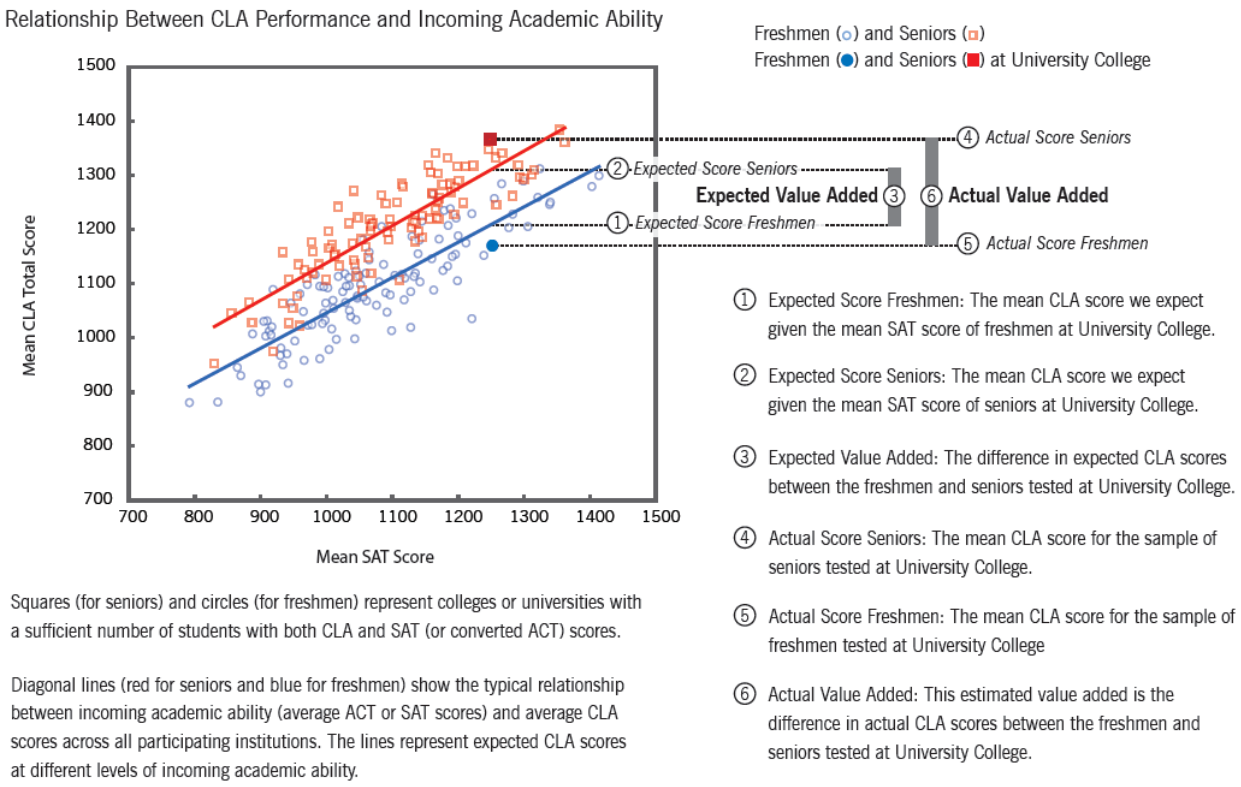

Figure 1. Graphical representation of the original CLA value-added score estimation approach.

those schools appear to have "grown" more in their critical thinking and writing skills than one would expect based on their entering academic ability.

Mathematically, the value-added score for school *j* is

$$VA_j = [\overline{CLA}_{j,sr} - E(\overline{CLA}_{j,sr})] - [\overline{CLA}_{j,fr} - E(\overline{CLA}_{j,fr})]$$

where $\overline{CLA}_{j,sr}$ and $\overline{CLA}_{j,fr}$ are the observed senior and freshman mean CLA scores at school *j*, and $E(\overline{CLA}_{j,sr})$ and $E(\overline{CLA}_{j,fr})$ are the corresponding expected values of those mean scores. The expected values, which are derived from the regression lines shown in Figure 1, reflect expected mean CLA performance given the average SAT of participating students. Simply put, a value-added score reflects the difference between the senior regression residual and the freshman regression residual. The formula can be rearranged as

$$VA_j = [\overline{CLA}_{j,sr} - \overline{CLA}_{j,fr}] - [E(\overline{CLA}_{j,sr}) - E(\overline{CLA}_{j,fr})]$$

to show that value-added scores indicate the difference between the observed and expected freshman-senior differences.

The New Approach

The HLM-based approach produces value-added scores that indicate the degree to which observed senior mean CLA scores exceed or fall below expectations established by two measures of entering academic ability: (1) mean SAT scores of the participating seniors and (2) mean freshman CLA scores. Although this approach does not employ difference scores (i.e., the average difference in performance between freshmen and seniors), it still provides scores that reflect average learning relative to expected. To illustrate, consider several schools admitting students with similar average performance on general academic ability tests (e.g., the SAT and ACT) and on tests of higher-order skills like critical thinking and written communication (e.g., the CLA). If, after four years of college education, the seniors at one school perform better on the CLA than is typical for schools admitting similar students, one can infer that more learning has taken place at the highest performing school. This is the basic idea behind value-added models based on “residuals” rather than “differences.”

As suggested earlier, a new value-added score estimation approach might obtain higher reliability and provide school-specific indicators of value-added score precision. The HLM-based approach attempts to realize both of these improvements. First, this approach may achieve gains in reliability because the school-level linear model specified in the HLM (from which value-added scored are derived) does not depend on difference scores. Second, this approach provides an estimate of value-added score precision for each school, which can be used to compute a unique 95% confidence interval for each school’s value-added score.

Statistically speaking, this approach incorporates two levels of analysis: (1) a student level for modeling CLA scores within schools as predicted by individual students' SAT scores and (2) a school level for modeling senior mean CLA scores as predicted by senior mean SAT and freshman mean CLA scores. A detailed statistical specification of this model is provided in the Appendix, but the basic ideas behind the computation of value-added scores can be explained in the terms of multiple regression. Consider an equation for predicting mean senior CLA scores from mean SAT scores of participating seniors and mean CLA scores of participating freshmen. In the HLM, this equation takes the form

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{SAT}_{\bullet j}) + \gamma_{02}(\overline{CLA}_{\bullet j,fr}) + u_{0j}$$

where β_{0j} is the mean CLA score at school j, $\overline{SAT}_{\bullet j}$ is the mean senior SAT score at school j, and $\overline{CLA}_{\bullet j,fr}$ is the mean freshman CLA score at school j. The γ coefficients are the school-level regression intercept and slope coefficients, and u_{0j} is the residual for school j. This residual, which reflect the difference between observed and expected mean senior CLA performance, serves as the value-added score.

Since they are based on different samples of students, it may not seem intuitive to include freshman mean CLA scores as predictors of senior mean CLA scores. However, analyses consistently indicate that freshman mean CLA adds significantly to the prediction of senior mean CLA scores. This finding indicates that freshman mean CLA and senior mean SAT capture somewhat different but nevertheless important characteristics of entering students' abilities. Furthermore, it can be argued that, in order for this to be accepted as a value-added model for CLA scores, there should be a control for entering CLA scores.

Methods

The remainder of this report addresses the following questions about the original and HLM-based value-added score estimation approaches:

1. Do the two approaches yield value-added scores that are similarly reliable and similarly consistent across years?
2. Do the two approaches yield comparable value-added estimates?
3. What additional information is provided by school-specific indicators of value-added score precision?

Most analyses were carried out using data gathered at 99 schools participating in the 2006-2007 CLA administration and 154 schools participating in the 2007-2008 CLA administration with at least 25 students participating. A subset of data from 71 schools participating in both administrations was used to study the consistency of value-added scores across years.

Reliability

Reliability for each approach was estimated using a modified version of the split-sample method described by Klein, Benjamin, Shavelson, and Bolus (2007). This method involves randomly splitting the freshman data gathered at each school into Samples A and B, doing the same for seniors, computing Sample A value-added scores and Sample B value-added scores, and then correlating the two sets of value-added scores. The modified approach included two improvements. First, a Spearman-Brown correction ($\frac{2r}{1+r}$, where r is the unadjusted split-half reliability) was used to adjust the reliability estimates for the use of half-size samples. This adjustment treats each school's value-added score as a composite of its Sample A and Sample B value-added scores and also treats the Sample A and Sample B value-added scores as parallel measurements, which seems reasonable because the samples were split randomly. Second, since

results depends on how the data were split, the average of a random sample of 1,000 split-sample reliabilities was computed in order to obtain a more stable estimate of reliability. Year-to-year value-added score consistency for both approaches was estimated by correlating value-added scores from consecutive CLA administrations.

Comparability

The first question a school might ask about a possible change in value-added score estimation is, "How would this change the value-added score for my school?" Indeed, critics point out that different models produce different results as one reason to distrust value-added scores (Banta & Pike, 2007). Correlations between value-added scores produced by the two approaches were computed to address the issue of comparability. Then, in order to estimate the "true" relationship, which is typically stronger than the observed relationship, the correlations were disattenuated for unreliability using the formula

$$\frac{r_{XY}}{\sqrt{r_{XX'}r_{YY'}}}$$

where r_{XY} is the observed correlation between the original and HLM value-added scores, $r_{XX'}$ is the reliability of the original value-added scores, and $r_{YY'}$ is the reliability of the HLM value-added scores (estimated using the approach described above).

A simulation study was also carried out to corroborate the disattenuated correlations. Specifically, this analysis attempted to simulate what CLA data (and subsequent value-added scores) would look like if all schools tested all freshmen and seniors (i.e., when there is no sampling error). In order to accomplish this, the distributions (sample sizes, means, standard deviations, and covariance matrices) of CLA and SAT scores within and between actual schools were estimated. These values were first used to create 200 simulated freshman classes consisting

of anywhere from a few hundred to a few thousand students. The true average gain between freshman and senior year was then simulated for these freshman classes and added to the freshman scores in order to obtain 200 simulated senior classes. Based on observed gains (adjusted for differences in ability), it was assumed that the distribution of average CLA gain scores had a mean of 85 CLA scale points (reflecting an effect size around 0.65) with a standard deviation of 45.

Selective attrition from the senior class was simulated to account for students dropping out of school. A logistic function was employed to calculate each student's probability of dropping out such that students with SAT scores 1.5 standard deviations below the mean in each school had a 0.50 probability of dropping out. This level of attrition produced simulated senior classes with average SAT scores about 23 points higher than their respective freshman classes, which is quite similar to average observed differences.

Simulated data from the full freshman classes and the senior classes with attrition were used to compute value-added scores using both approaches. Finally, correlations between those value-added scores were computed.

Indicators of Precision

The new method for estimating value-added, which employs HLM, provides school-specific indicators of value-added precision. Given that value-added scores are not perfectly reliable, it is prudent to condition interpretations of these scores on available information about their precision (or lack thereof). In order to demonstrate how the HLM-based approach provides additional information about value-added score precision, 95% confidence intervals were generated using the standard errors provided by the HLM estimation. Value-added scores and accompanying 95% confidence interval were plotted together. In addition, the relationship

between value-added score precision and sample size was examined by plotting the standard errors versus the number of seniors tested in each school.

Results

Reliability and Consistency

Analyses indicate that the HLM-based approach produces more reliable value-added scores than the original approach. For the original approach, the average split-sample reliability was 0.730 in 2006-2007 and 0.635 in 2007-2008, which is slightly higher than the value of 0.63 reported by Klein et al. (2007). The corresponding average reliabilities were 0.809 and 0.749 for the new approach.

Year-to-year consistency, as indicated by the correlation between value-added scores in adjacent test administrations, was computed for 71 schools participating in the 2006-2007 and 2007-2008 CLA administrations. The increase in year-to-year consistency from the original approach (0.320) to the new approach (0.583) indicates that value-added scores from the new approach are dramatically more stable across years.

To further illustrate the gain in consistency, 40 schools that participated in 3 recent CLA administrations (2005-2006, 2006-2007, and 2007-2008) were divided into three groups based on value-added score consistency (Figure 1): low (range of three value-added scores greater than 1.5), moderate (range between 0.75 and 1.5), and high (range less than 0.75). With the original model, 40% of schools had low consistency, 38% had moderate consistency, and 23% had high consistency. Using the HLM model, the percentage with moderate consistency increased to 55%, and the percentage with low consistency decreased to 25%. The percentage with high consistency was about the same (20%).

Figure 1. Examples of low, moderate, and high value-added score consistency across three CLA administrations (value-added estimates shown as dots, 95% confidence intervals shown as solid lines).

Comparability

Correlations between the value-added scores produced by the two approaches were 0.799 and 0.718 in the 2006-2007 and the 2007-2008 data sets, respectively. These correlations indicate that the two approaches produce similar but far from identical results (Figure 2).

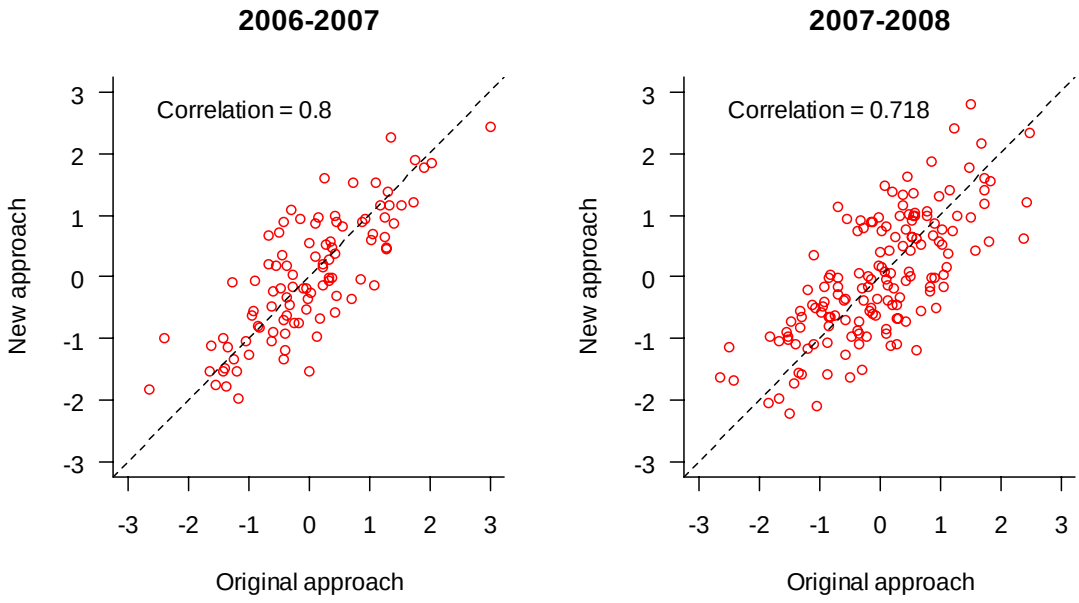

Figure 2. Scatterplots showing value-added scores produced by the original and new estimation approaches.

To provide a clearer representation of the true underlying relationship, the correlations were disattenuated using the reliability estimates presented in the previous section. The rounded disattenuated correlations were 1.00 for the 2006-2007 data and 1.00 for the 2007-2008 data.[2] This finding suggests that the relative standing of schools would be identical regardless of the value-added model if the value-added scores were perfectly reliable. In short, it is likely that both approaches would yield essentially the same results if all schools tested much larger samples of students.

This result was corroborated by the simulated data analysis. The correlation between value-added scores from the two models for the simulated data was 1.00, which indicates that, although the two value-added models produce differing results, those differences are fully accounted for by the unreliability of the value-added scores. Thus, it seems that the value-added scores generated by the two models are estimates of the same underlying construct: learning relative to expectations.

Indicators of Precision

Given that value-added scores are not perfectly reliable, it is prudent to condition interpretations of these scores on available information about their precision (or lack thereof). Unlike the original value-added estimation approach, the HLM-based approach provides school-specific indicators of value-added precision. To illustrate, Figure 3 shows 95% confidence intervals drawn as vertical bars above and below the value-added score point estimates for 2007-2008 (ordered from least to greatest value-added score). As an example, consider the school with the lowest value-added score (the leftmost point in Figure 3). With a value-added score of -2.2 on a standardized scale, seniors at this school scored roughly 75 CLA scale score points below expected. Given that the confidence interval for this school does not intersect with the horizontal

[2] The disattenuated correlations slightly exceeded 1.00.

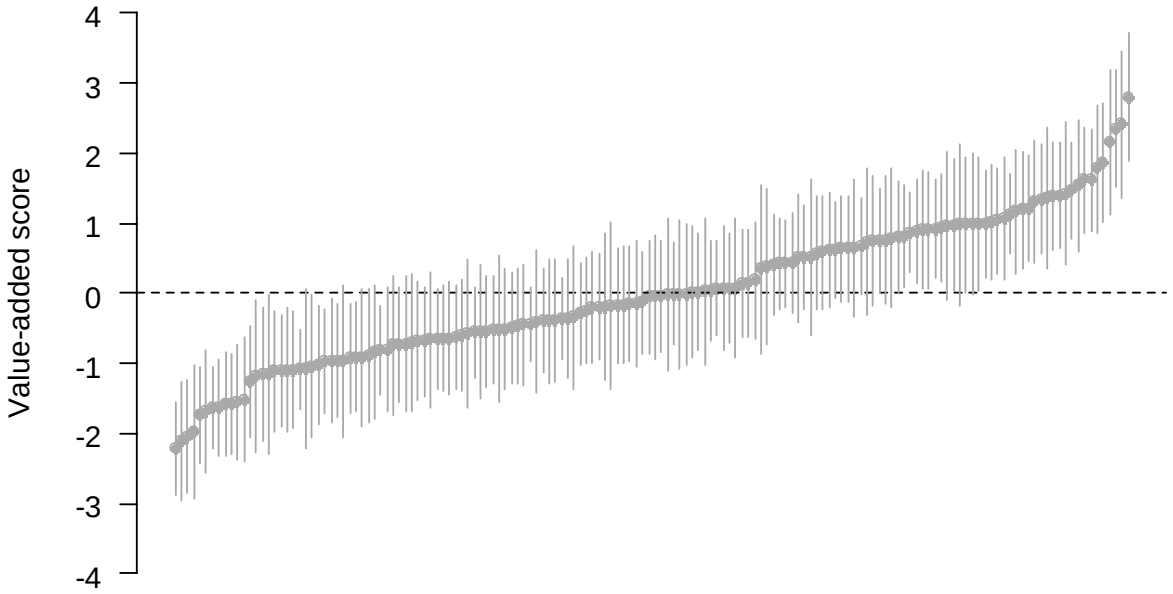

Figure 3. Value-added scores and 95% confidence intervals for the 2007-2008 data.

dashed line at 0, which reflects the "at expected" condition, one can conclude that this school has a value-added score that is significantly "below expected." Schools with confidence intervals that cross the 0 line could be classified as "near expected," and those with confidence intervals fully above the 0 line could be classified as "above expected." When schools have confidence intervals that overlap substantially, this raises uncertainty about the magnitude of the difference between those schools' value-added scores because the difference may reflect sampling error.

The confidence intervals shown in Figure 3 vary in size, and this primarily reflects differences between schools in the number of students taking the CLA. Schools testing a larger number of senior students will obtain more precise value-added estimates, which improves interpretability by providing greater confidence in the value-added point estimate and by making it easier to discern significant difference among value-added scores. Figure 4 shows that the size of the 95% confidence interval decreases sharply as sample size increases toward 100 students.

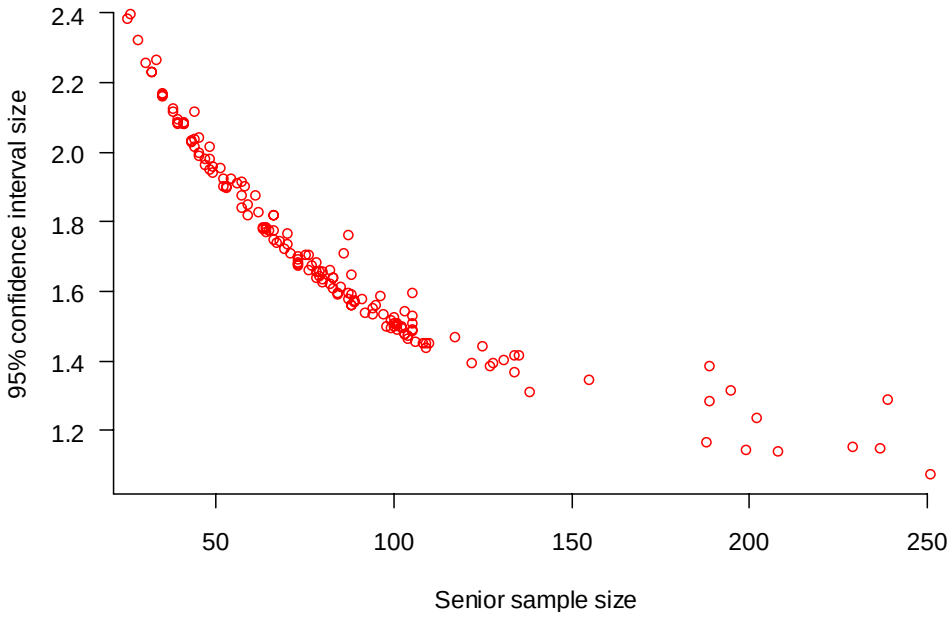

Figure 4. Relationship between 95% confidence interval size and senior sample size for the 2007-2008 data.

Discussion and Conclusions

This report presented two approaches to estimating institutional value-added scores for the CLA. A comparison of these methods revealed that they produce highly correlated value-added scores and that they would produce virtually identical value-added scores if sampling error was eliminated. Given this fact, one should prefer the estimation approach that generates the most reliable value-added scores for a given number of students tested. The proposed HLM-based approach is more efficient (cost effective) in the sense that, when the number of students tested is held constant, scores from the new approach are more precise within a year and are more realistically stable across years. In addition, the new approach provides school-specific indicators of value-added score precision, which improve the interpretability of scores.

The statistical techniques involved in the new approach have been available for several decades, but this represents a first effort to employ these techniques to estimate value-added

scores for institutions of higher education. Thus, the research staff working on the CLA believes that this work reflects advancement for the nascent field of institutional value-added estimation and for the quality and interpretability of CLA scores. For these reasons, CLA value-added scores will be estimated using the new approach described here starting in the 2009-2010 administration cycle.

References

Banta, T. W., & Pike, G. R. (2007). Revisiting the blind alley of value added. *Assessment Update, 19*(1), 1-2, 14-15.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Holt, Rinehart and Winston, Inc.

Keller, C. M., & Hammang, J. M. (2008). The voluntary system of accountability for accountability and institutional assessment. *New Directions for Institutional Research, 2008*(S1), 39-48.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415-439.

Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review, 32*(6), 511-525.

Appendix: Detailed specification of HLM-based value-added approach

Student level

$$CLA_{ij} = \beta_{0j} + \beta_{1j}(SAT_{ij} - \overline{SAT}_{\bullet j}) + r_{ij}$$

CLA_{ij} is the CLA score of student *i* at school *j*.
SAT_{ij} is the SAT score of student *i* at school *j*.
$\overline{SAT}_{\bullet j}$ is the mean SAT score at school *j*.
β_{0j} is the mean CLA score at school *j*.
β_{1j} is the within-school CLA-on-SAT regression slope at school *j*.
r_{ij} is the residual for student *i* in school *j*; $r_{ij} \sim N(0, \sigma^2)$.

School level

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{SAT}_{\bullet j}) + \gamma_{02}(\overline{CLA}_{\bullet j,fr}) + u_{0j}$$

$\beta_{1j} = \gamma_{10}$ (assuming that all schools have the same within-school CLA-on-SAT regression slope)

$\overline{CLA}_{\bullet j,fr}$ is the mean freshman CLA score at school *j*.
γ_{00} is the intercept of the school-level value-added equation.
γ_{01} is the school-level slope for predicting senior mean CLA from senior mean SAT.
γ_{02} is the school-level slope for predicting senior mean CLA from freshman mean CLA.
γ_{10} is the pooled within-school CLA-on-SAT slope.
u_{0j} is the residual for school *j* (i.e., value-added score); $u_{0j} \sim N(0, \tau)$.

Mixed model (combination of school- and student-level equations)

$$Y_{ij} = \gamma_{00} + \gamma_{01}(\overline{SAT}_{\bullet j}) + \gamma_{02}(\overline{CLA}_{\bullet j,fr}) + \gamma_{10}(SAT_{ij} - \overline{SAT}_{\bullet j}) + u_{0j} + r_{ij}$$