

WHITE PAPER

ADVANCING INSTITUTIONAL VALUE-ADDED SCORE ESTIMATION

JEFFREY STEEDLE

DRAFT: June 1, 2009



collegiatelearningassessment
COUNCIL FOR AID TO EDUCATION

215 Lexington Avenue, Floor 21 New York NY 10016-6023
p | 212.217.0700 f | 212.661.9766 e | cla@cae.org w | www.cae.org/cla

Abstract

For the purposes of institutional assessment, many post-secondary schools administer the Collegiate Learning Assessment (CLA) in order to obtain value-added scores, which indicate whether average gains in critical thinking and writing skills are below, near, or above what one would expect at schools with similar entering academic ability. This report presents analyses comparing the statistical qualities of value-added scores generated by the original CLA value-added approach and a new approach that employs hierarchical linear modeling (HLM) and a different equation for computing expected mean CLA scores. Value-added scores produced by the two approaches are highly correlated and would be essentially identical if large samples of students were available at all schools. Reliability analyses reveal that the new approach produces value-added scores that are slightly more reliable within years and substantially more reliable across years than those generated by the original approach. As an added benefit, the HLM-based approach generates school-specific indicators of value-added score precision, which reflect variation in precision across schools (i.e., schools with larger sample sizes get greater precision). For these reasons, the new HLM-based approach is viewed as an improvement over the original approach and will therefore be implemented to estimate CLA value-added scores starting in the 2009-2010 assessment cycle. Schools administering the CLA should expect to observe greater overall stability in value-added scores across years, and they will receive school-specific indicators of value-added score precision, which help inform interpretations of value-added scores.

Institutional value-added scores intend to capture whether the growth in achievement between freshman and senior year at a given school is below, near, or above what is typically observed at schools with students of similar entering academic ability. The estimation of value-added scores for institutions of higher education is a relatively new enterprise, so few studies have evaluated or compared the statistical properties of alternative institutional value-added score estimation approaches. This report presents such analyses using data collected over the course of recent administrations of the Collegiate Learning Assessment (CLA), which is a measure of college students' critical thinking and writing skills as applied to authentic, open-ended problems (www.cae.org).

Since the initial administration of the CLA during the 2004-2005 school year, its value-added scores have been estimated using an ordinary least squares (OLS) regression approach that involves subtracting a freshman class residual score from a senior class residual score (described in greater detail below). Value-added scores are estimated using samples of students because census testing (i.e., testing all freshmen and all seniors) is logistically unfeasible and prohibitively expensive. Typically, a school administers the CLA to roughly 100 freshmen during the fall and 100 seniors during the following spring of the same academic year. Prior research demonstrated that value-added scores generated by this approach are reliable in a given year (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 current CLA value-added approach should provide reasonably accurate estimates of the schools' true value-added scores.

That said, one could imagine a value-added estimation approach producing scores with even greater reliability, which corresponds to greater precision (or equivalently less error). Current value-added scores are based on differences between freshman and senior residual scores, but difference scores are less reliable than the scores from which they are derived, especially when those scores are correlated (Crocker & Algina, 1986). When the school is the unit of analysis, freshman and senior residual scores typically correlate about 0.40, which helps explain why value-added scores are less reliable than freshman or senior residual scores alone (Klein et al., 2007).

A value-added estimation approach, then, that does not depend on difference scores may provide better within-year value-added score reliability. Moreover, an increase in within-year reliability would likely increase the year-to-year reliability of value-added scores. This does not mean, of course, that value-added scores should be identical across years (e.g., due to sampling error, programmatic improvements, or other measurement error), but they should not change radically either. Substantial value-added score variability over time diminishes the ability to interpret differences in these scores 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 current method of estimating value-added scores 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 value-added score precision because precision varies from one school to another depending, for example, on the size of the sample of students participating in the CLA. Acknowledging this uncertainty would facilitate a school's interpretation of its own value-added score by providing a realistic sense of the possible range that school's score might take if it had been based on other samples of students who might have participated in the CLA that year (e.g., a 95% confidence interval). Information about score precision could also be used to determine what would constitute credible and trustworthy

differences in value-added scores between schools. Precision improves as sample size increases, so schools would be rewarded with greater precision (i.e., narrower confidence intervals) and therefore greater interpretability of value-added results by meeting or exceeding sample size targets.

As part of continual efforts to improve the CLA, analyses were carried out to compare the original approach the CLA program employs to estimate value-added scores with a new approach that could provide the improvements described above. The following sections provide background on these approaches and the results of analyses comparing their statistical properties.

The Original Approach

Although CLA data are most often collected using a cross-sectional design (i.e., freshmen and seniors in the same academic year), average differences between freshmen and seniors are often interpreted as indicators of “growth” in critical thinking and writing skills from freshman to senior year. At nearly all schools, seniors outperform freshmen on average, but the average CLA freshman-senior difference varies widely across schools. For a given school, the value-added score produced by the original estimation approach indicates the degree to which the observed freshman-senior difference is below, near, or above expectations, where expectations reflect the freshman-senior difference one typically sees at other schools with comparable students (as measured by SAT or ACT scores). Schools at which the freshman-senior differences exceed expectations are said to have high “value added” because students attending those schools appear to have “grown” more in their critical thinking and writing skills than students at other schools after controlling for entering academic ability.

Statistically speaking, a school’s CLA value-added score equals the school’s senior residual score minus its freshman residual score. (This process is depicted in Figure 1 for a fictional school called University College.) Obtaining freshman residual scores involves the use of OLS regression to estimate the linear relationship between average entering academic ability and average CLA performance for freshmen. The OLS residuals from this school-level analysis serve as the freshmen residual scores, which reflect average CLA performance relative to the performance one would expect from a group of students with entering academic ability similar to that of students taking the CLA. A corresponding procedure is used to obtain school-level residuals for seniors (i.e., mean CLA scores of seniors are regressed on their mean SAT scores). Value-added scores are obtained by subtracting the freshman residual score from the senior residual score.¹

¹ The difference between residual scores is mathematically equivalent to the difference between the observed freshman-senior difference and the expected freshman-senior difference.

$$\begin{aligned}
 & \text{Senior Residual} - \text{Freshman Residual} \\
 &= (\text{Senior}_{\text{Observed}} - \text{Senior}_{\text{Expected}}) - (\text{Freshman}_{\text{Observed}} - \text{Freshman}_{\text{Expected}}) \\
 &= (\text{Senior}_{\text{Observed}} - \text{Freshman}_{\text{Observed}}) - (\text{Senior}_{\text{Expected}} - \text{Freshman}_{\text{Expected}}) \\
 &= \text{Observed freshman-senior difference} - \text{Expected freshman-senior difference}
 \end{aligned}$$

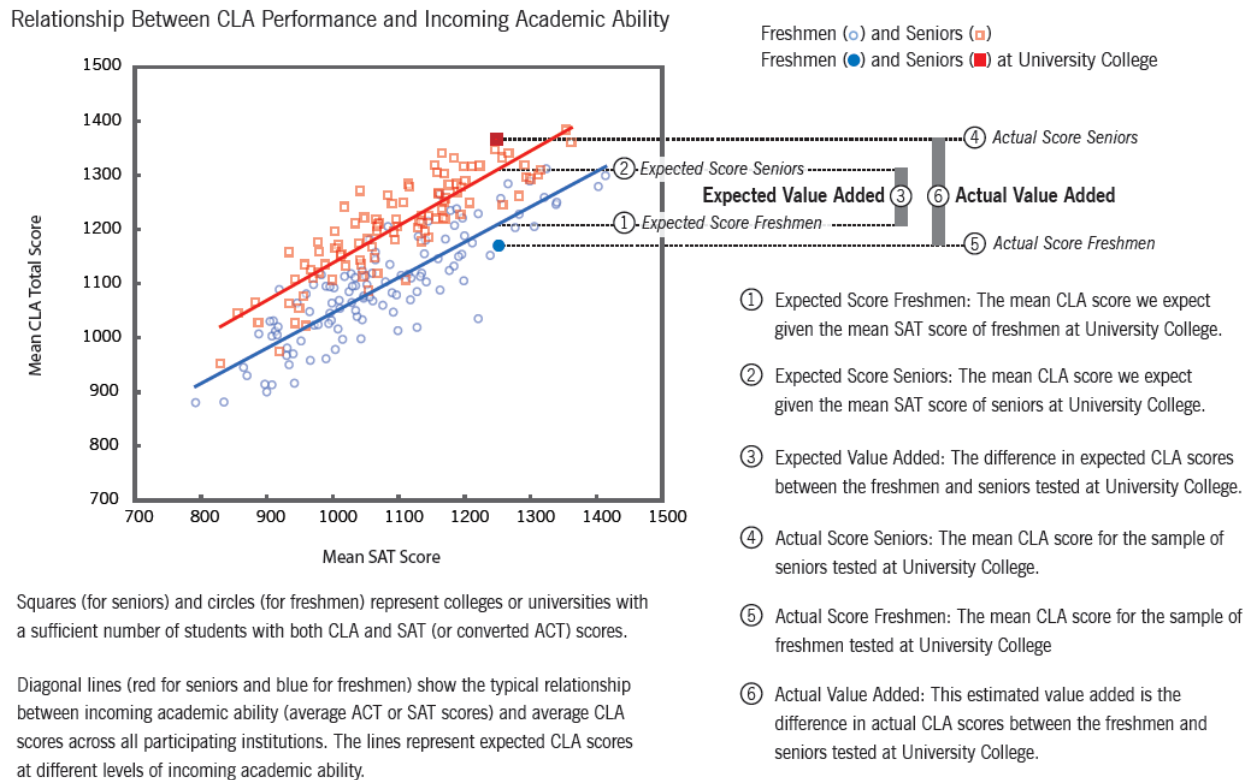


Figure 1. Graphical representation of the current value-added score estimation approach.

The New Approach

An alternative approach to estimating institutional value-added scores involves comparing the CLA performance of seniors at one school to the CLA performance of seniors at other schools admitting students with similar academic skills. To illustrate, consider a group of similarly selective schools admitting students with comparable academic skills (critical thinking and writing skills in addition to more general academic skills measured by the SAT or ACT). If, after four years of education, the seniors at one school demonstrate superior CLA performance relative to the other schools in the group, one could infer that students at the high performing school tend to “grow” more in their critical thinking and writing skills than students at the other schools in the group.

This is the reasoning behind the new value-added estimation approach presented here. This approach produces school-level 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 (or ACT) scores of the participating seniors and (2) mean freshman CLA scores. Both measures are correlated with the senior mean CLA scores, and both are statistically significant predictors of senior mean CLA scores when entered simultaneously in a regression equation. These findings indicate that the CLA and SAT capture somewhat different but nevertheless important characteristics of entering students’ abilities.

As suggested earlier, an alternative value-added score estimation approach might obtain higher reliability by not using difference scores and might provide school-specific indicators of value-added score precision. The new approach, which employs hierarchical linear modeling (HLM), attempts to realize both of these improvements. First, the school-level linear model specified in the HLM (from which value-added scores are derived) does not depend on difference scores. Second,

this approach provides estimates of the precision of senior mean CLA scores and value-added scores at each school. This multilevel 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 (and for estimating institutional value-added scores).

Although HLM is employed to account for CLA score variation within and between schools, value-added scores can be explained in the terms of OLS multiple regression. From this perspective, the senior mean CLA score is the outcome variable, and the senior mean SAT score and that year's freshman mean CLA score are predictor variables. A school's residual score indicates the difference between the observed senior mean CLA score and the expected senior mean CLA score (a linear combination of the mean SAT score of participating seniors and the mean CLA score of participating freshmen). This residual serves as the value-added score for a school. HLM provides an estimate of the precision of the residual (the posterior variance of the school-level residual), which can be used to compute a 95% confidence interval for the value-added score.

Comparing the Alternatives

In order to inform the decision of which value-added score estimation approach to use in future CLA administrations, the remainder of this report addresses the following questions about the original and new approaches:

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

Unless otherwise noted, 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. Data gathered at 71 schools participating in both administrations were used to study the stability of value-added scores across years.

Comparability

The first question a school might ask about a possible change in value-added score estimation is, "Would the value-added score for my school change?" In other words, "Will my school's value-added estimate be better or worse with the new approach compared to the original one?" 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).

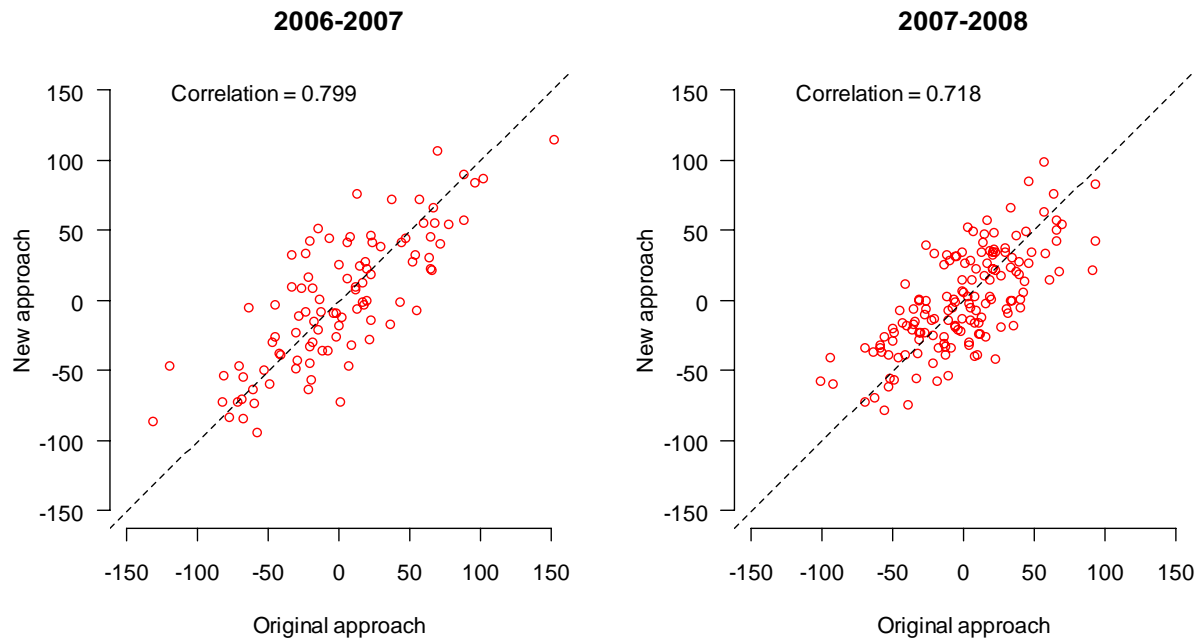


Figure 2. Scatterplots showing value-added scores produced by the original and new estimation approaches.

These correlations reflect the *observed* relationships between value-added scores, which may not be as strong as the *true* relationships because of unreliability. To provide a clearer representation of the true underlying relationship, the correlations were disattenuated using reliability estimates presented in the following 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 two approaches would rank order schools identically if value-added scores were perfectly reliable. The strength of the true relationship was corroborated using simulated CLA data, which reflected what CLA data would look like if all participating schools tested *all* freshmen and seniors. These simulated mean CLA scores reflect the simulated true scores at the schools because sampling error is eliminated. The correlations for the simulated data were 1.000 and 0.995 for the 2006-2007 and 2007-2008 simulated data, respectively. These results indicate that the observed differences between value-added scores produced by the two approaches stem from the less-than-perfect reliability of the value-added estimates. In short, it is likely that both approaches would yield essentially the same results if there were much larger samples of students.

Reliability

The reliability of CLA value-added scores was first reported in Klein, Benjamin, Shavelson, and Bolus (2007). They applied a novel split-sample approach to evaluate CLA score reliability. This involved 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. Using data from 44 schools participating in the 2005-2006 CLA administration, they estimated the reliability of value-added scores as 0.63. As

² The disattenuated correlations slightly exceeded 1.00.

the authors noted, this value is overly conservative due to the use of half-size samples. In addition, their results were based on one of a very large number of possible random splits, which means that the results could have come out differently if other random splits were studied.

For this report, a modified split-sample reliability estimation approach was applied to the 2006-2007 and 2007-2008 CLA data. 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 (treating a school’s value-added score as a composite of the Sample A and Sample B value-added scores and treating a school’s Sample A and Sample B value-added scores as parallel measurements). Second, the mean of a random sample of 1,000 split-sample reliabilities was computed in order to obtain a stable estimate of the expected value of reliability.

Analyses of data from both test administrations indicate that the new HLM approach produces more reliable value-added scores than the original approach (Table 1). For the original approach, the mean split-sample value-added reliability was 0.730 in 2006-2007 and 0.635 in 2007-2008. The corresponding mean reliabilities were 0.809 and 0.749 for the new approach.

Table 1.
Mean adjusted split-sample estimates of the within- and between-year reliability of value-added scores with the original and new method for estimating value-added scores

| Sample | Original | New |
|--------------|----------|-------|
| 2006-2007 | 0.730 | 0.809 |
| 2007-2008 | 0.635 | 0.749 |
| Year-to-Year | 0.320 | 0.583 |

As discussed earlier, year-to-year stability should also be of interest when evaluating the qualities of value-added score estimates. The same data used to estimate within-year split-sample reliability were used to estimate year-to-year split-sample reliability for 71 schools participating in the 2006-2007 and 2007-2008 CLA administrations. Results provided in Table 1 indicate that value added scores from the new approach are dramatically more reliable across years than scores from the original approach.

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). To illustrate, Figure 3 shows 95% confidence intervals drawn as vertical bars above and below the point estimates of value-added for schools in the 2007-2008 data (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). On average, 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 dashed line at 0, which reflects the “at expected” condition, one can conclude that this school has a value-added score that is “below expected” in a statistically significant way. 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.

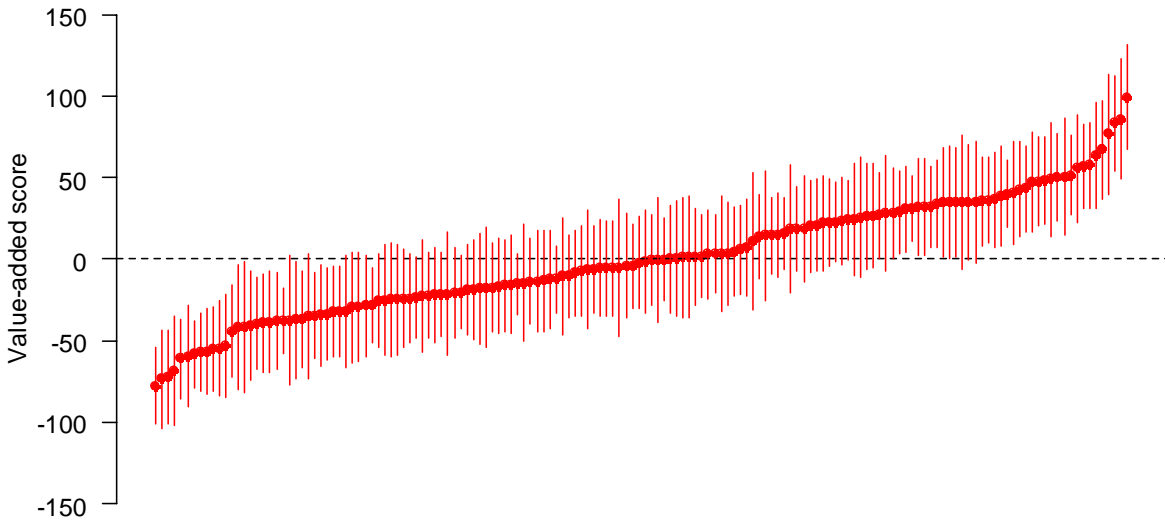


Figure 3. Value-added scores and 95% confidence intervals for the 2007-2008 data.

The confidence intervals shown in Figure 3 vary in size across schools, and this primarily reflects differences between schools in the number of students taking the CLA. Schools testing a larger number of students will obtain more precise value-added estimates (i.e., smaller confidence intervals), which improves the interpretability of value-added scores. Figure 4 shows that the size of the 95% confidence interval decreases sharply as sample size increases toward 100 students. Precision continues to increase beyond 100 students, but at a slower rate.

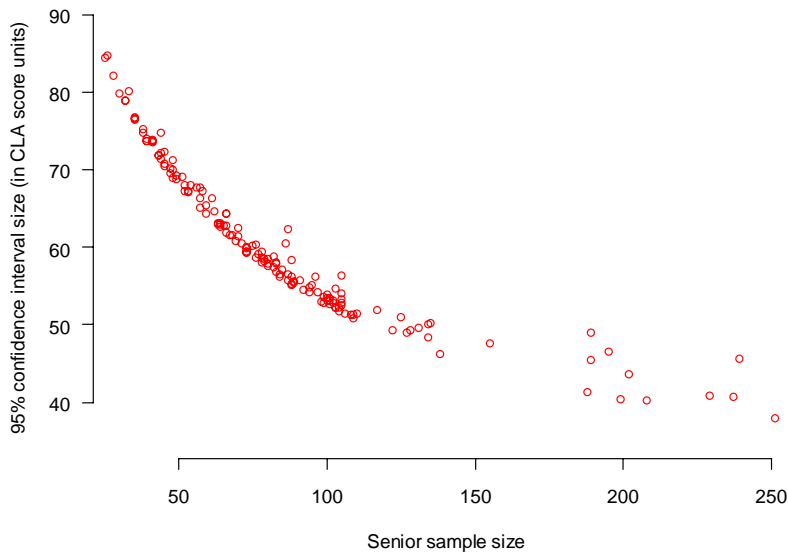


Figure 4. Relationship between 95% confidence interval size and senior sample size for the 2007-2008 data.

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

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