The Lumina Longitudinal Study: Summary of Procedures and Findings

Preface

The Collegiate Learning Assessment (CLA) tests the critical thinking, analytic reasoning, problem solving, and writing skills of college and university students using two types of tasks: Performance and Analytic Writing. Schools use CLA results to gauge how well their students, as a group, perform relative to the sample of all other schools administering the CLA and to a more select sample of schools admitting students of similar entering academic ability (as measured by the mean SAT or ACT scores of participating students). Additionally, the CLA examines whether the average difference between a school's freshmen and seniors is larger or smaller than it is at other schools with students of similar entering academic ability. In other words, are the students at a particular school improving as much as students at other schools? This type of *cross-sectional* comparison provides an estimate of the results that would be obtained from a *longitudinal* analysis (i.e., an analysis in which schools test entering freshmen and then retest those same students four years later when most of them are graduating seniors).

Purpose

As discussed in previous articles (Klein et al., 2007, 2008), there are arguments for both the cross-sectional and longitudinal approaches to CLA data collection. To help inform this debate with empirical evidence, the Lumina Foundation supported a five-year study that explored the feasibility and consequences of using each model.[1] Specifically, this study sought answers to the following questions:

- Did the cross-sectional and longitudinal analyses yield similar estimates of freshman-to-senior gain scores? And, did the answer to this question depend on which type of CLA task was studied (Performance Task or Analytic Writing Task)?
- Did most schools have similar freshman-to-senior gain scores, or were there large differences in gain scores among schools?
- Was the reliability of the cross-sectional CLA gain scores comparable to the reliability of the longitudinal gain scores?

The longitudinal feature of the Lumina study and its three phases of testing also allowed us to investigate two other matters related to CLA scores, namely:

- Was the improvement in scores between entering freshmen and "rising juniors" (i.e., students completing their sophomore year) comparable to the improvement between

[1] Our proposal also promised to shed light on the performance of minorities on the CLA. With that in mind we recruited a number of universities with substantial numbers of African-American and Hispanic students. The result is one of the largest data sets created to date that permits extensive comparative analysis of minority versus majority students' value-added growth in student learning as measured by the CLA. Early in the development of the Lumina Longitudinal study, CAE formed a partnership with Richard Arum, a professor of sociology and education at New York University and a researcher at the Social Science Research Council (SSRC) to examine the patterns of racial/ethnic inequality in the Lumina longitudinal CLA data set. We endorsed Dr. Arum's request to the Lumina and Ford Foundation for funding to identify a broad set of individual, social, and institutional factors associated with learning in higher education. In a preliminary report from this SSRC project, *Richard Arum and Josipa Roksa with Melissa Velez (2008), Learning to Reason and Communicate in College: Initial Report of Findings from the CLA Longitudinal Study, New York*, SSRC presents a number of noteworthy conclusions and recommendations about the student learning attained by socially disadvantaged groups in college. Richard Arum and Josipa Roksa now have produced a full length book manuscript forthcoming, Chicago: University of Chicago Press, fall 2010. We are pleased that this important work on the effects of social inequality on student learning has been produced by these authors associated with the SSRC.

rising juniors and graduating seniors? And, did the answer to this question depend on which type of CLA task was studied?

- How well do SAT and freshman CLA scores, by themselves and in combination, predict a student's college GPA?

Overview

We begin by summarizing the major features of the CLA program and its measures. Next, we describe the cross-sectional and longitudinal designs we employed, highlight how the Lumina longitudinal students differed from those in the larger population of students in the CLA program's database, and report the number of longitudinal schools and students that participated as freshmen and again as rising juniors and seniors. Finally, we discuss our findings and conclusions regarding the questions listed above and their implications.

CLA Measures

Unlike most standardized tests in postsecondary education, the CLA does not include any multiple-choice or true/false questions. Instead, the CLA program uses several types of "constructed response" tasks (i.e., students create their own answers like an essay test). In a typical administration of the CLA, students complete either a *Performance Task* or an *Analytic Writing Task*. Students participating in the Lumina Longitudinal Study completed both. For a complete description of CLA tasks and samples of student performance, see *Architecture of the CLA Tasks* (CAE, 2009).

In the Performance Task, students are instructed to draft a letter, memo, or similar document (e.g., to a supervisor, co-worker, or company) about some matter. They also are given a "library" of documents, some of which are more credible and relevant to the task than others, and some may include contradictory information. Students have 90 minutes to evaluate the information provided, synthesize and organize that evidence, draw conclusions, and create a cogent response.

The Analytic Writing Task consists of two sections. First, students are allotted 45 minutes for the *Make-an-Argument* task in which they present their perspective on an issue like "Government funding would be better spent on preventing crime than dealing with criminals after the fact." Next, the *Critique-an-Argument* task gives students 30 minutes to identify and describe logical flaws in an argument. Here is one example:

> Butter has now been replaced by margarine in Happy Pancake House restaurants throughout the southwestern United States. Only about 2 percent of customers have complained, indicating that 98 people out of 100 are happy with the change. Furthermore, many servers have reported that a number of customers who still ask for butter do not complain when they are given margarine instead. Clearly, either these customers cannot distinguish margarine from butter, or they use the term "butter" to refer to either butter or margarine. Thus, to avoid the expense of purchasing butter, the Happy Pancake House should extend this cost-saving change to its restaurants in the southeast and northeast as well.

Procedures

The CLA item bank contains several prompts of each type described above. Students who participated in all three phases of Lumina testing took one prompt of each type at baseline as freshmen in fall 2005, again as so-called "rising juniors" in the spring of 2007, and once more as seniors in spring 2009. Prompts were assigned randomly to students within schools. Thus, students sitting near one another were usually answering different prompts. Students did not repeat prompts they took on previous testing occasions.

Tasks were timed separately and administered by computer under proctored conditions at each school during a multi-week testing window. Students always took a Performance Task before they took the Analytic Writing Task (Make-an-Argument and Critique-an-Argument).

Performance Task responses were scored by trained readers, and responses to the Analytic Writing Tasks were machine scored (see Klein, 2007).

To facilitate comparisons among measures (and to allow combining scores across them), the human- or machine-assigned raw scores on each task were converted to a score distribution that had the same mean and standard deviation as the SAT total scores of the population of freshmen CLA takers who took that task. The seniors' raw scores were converted to scale scores using the same formulas used with freshmen so that any differences in answer quality between classes would not be obscured by the score conversion process.

A school's (and a student's) total scale score was the sum of its weighted Performance Task, Make-an-Argument, and Critique-an-Argument scale scores, where the weights were 50 percent, 25 percent, and 25 percent, respectively. A school's Analytic Writing score was the average of its Make-an-Argument and Critique-an-Argument scale scores.

Participation and Attrition

At the start of the Lumina longitudinal study, 50 schools agreed to test approximately 300 of their entering freshmen in fall 2005 and participate in subsequent phases of testing. The Lumina schools had characteristics that were similar to the non-Lumina schools in the CLA database. For example, the percentage of enrollees who were Black in the Lumina schools and the CLA database were 17 percent and 12 percent, respectively. The analogous numbers for Hispanic students were 8 percent and 7 percent. In both data sets, 50 percent of institutions were public. The only notable difference was that the Lumina schools were somewhat more selective than other schools in the CLA database.

A student was classified as having participated in a phase of testing if that student completed at least one of the three CLA tasks in that phase; a school was classified as having participated in a phase if it had at least 25 students participating in that phase.

A total of 9,167 Lumina freshmen completed the fall 2005 testing, but only 1,330 of them (14 percent) eventually completed all three phases of testing. Most of the attrition was due to schools rather than individual students dropping out of the study (although some schools may have dropped out because of difficulty recruiting students to participate). Only 26 (52 percent) of the initial 50 schools tested at least 25 students in both Phases 1 and 3; just 20 schools (40 percent) met the minimum sample size requirements in all three phases. These 20 schools tested 4,748 freshmen in the fall of 2005, 2,327 rising juniors in the spring of 2007, and 1,675 seniors in the spring of 2009.

On the average, a school that stayed in the study for all three phases lost about one-third of its students that participated as freshmen. Thus, although this is a substantial loss, it is far less than the overall attrition rate. We found that dropouts were more likely to be Black or Hispanic, non-native English speakers, and students with total SAT scores about 80 points lower than their classmates. However, even when taken together, these student-level characteristics explained only five percent of the variance in students' decisions to drop out of the study (but perhaps not from the school). We looked for but did not find any school characteristics associated with dropping out of the study.

Table 1 shows the number of schools that tested at least 25 students in the indicated phase or combination of phases. For this table, "participation" was defined as completing all three types of CLA measures. The last column shows the number of students participating at the schools shown in the middle column. For example, there were 26 schools with at least 25 students participating in the fall 2005 and spring 2009 phases of testing; there were 2,049 students at these 26 schools.

Table 1
Number of schools and students participating in the Lumina Longitudinal Study

Phase(s)	Schools	Students
1 (Fall 2005)	47	9,167
2 (Spring 2007)	32	3,137
3 (Spring 2009)	30	2,289
Both 1 and 2	29	2,860
Both 1 and 3	26	2,049
Both 2 and 3	20	1,409
All 3	20	1,330

Most of the schools participating in this study also tested a sample of their spring 2006 graduating seniors. Testing seniors in spring 2006 made it possible to compare their performance (“cross-sectionally”) with the performance of the freshmen in the longitudinal sample who tested in fall 2005. However, unlike the students in the longitudinal cohort, the cross-sectional seniors took either one Performance Task or one Make-an-Argument Task plus one Critique-an-Argument Task, rather than all three measures. There were 25 schools with at least 25 spring 2006 seniors in this cross-sectional group and a total of 2,615 seniors at these schools who completed at least one of the three tasks.

Results

We now turn to the questions that were asked at the beginning of this report. Unless otherwise noted, the analyses treat the school as the unit of analysis, which makes school average scores the primary data. Average differences between freshmen and seniors are converted to “effect sizes” by dividing those differences by their respective freshman CLA score standard deviations. This step puts the score differences on a common scale, which allows for making direct comparisons among measures.

Did the cross-sectional and longitudinal analyses yield similar estimates of freshman-to-senior gain scores? And, did the answer to this question depend on which type of CLA task was studied?

To answer this question, we computed cross-sectional effect sizes for each school by subtracting the mean fall 2005 freshman score from the mean spring 2006 senior score, and then dividing the result by the freshman standard deviation. The longitudinal effect size was computed by subtracting the mean fall 2005 freshman score from the mean score of these same students as seniors in spring 2009, and then dividing the result by their freshman standard deviation. Note that both calculations used the same set of freshmen.

Table 2 shows that, on average across participating schools, the Performance Task effect sizes estimated using cross-sectional data were similar to those obtained using the Phase 1 to 3 longitudinal data. Specifically, the average four-year effect sizes for Performance Task scores with the cross-sectional and longitudinal methods were 0.5 and 0.4, respectively. In contrast, the cross-sectional effect sizes for the Analytic Writing Tasks were substantially larger than the longitudinal effect sizes for these tasks. Students participating in the longitudinal study underperformed on the Analytic Writing Tasks relative to cross-sectional seniors taking these same tasks, which inflated the cross-sectional effect sizes (i.e., making the differences between freshmen and seniors larger than it should be). We suspect this was due to a combination of fatigue and reduced motivation as a result of the longitudinal students taking a 90-minute Performance Task immediately before taking the two Analytic Writing Tasks. In contrast, spring

2006 seniors took either one Performance Task or a combination of one Make-an-Argument and one Critique-an-Argument Task.

Table 2
Average cross-sectional and longitudinal effect sizes

Score	Cross-sectional	Longitudinal
Performance Task	0.5	0.4
Analytic Writing Task	1.2	0.7
Total	0.9	0.7

Note: The cross-sectional effect sizes reported here are not adjusted for average differences in entering academic ability between the fall 2005 freshmen and the spring 2006 seniors. While these sorts of adjustment are possible, their accuracy has yet to be demonstrated.

The correlations between the cross-sectional and longitudinal freshman-to-senior effect sizes for the Performance Task, Analytic Writing Task, and total scores were also computed. These correlations were .47, .42, and .49, respectively, but these values were biased downwards by a few outliers. For example, when we excluded the two largest outliers on the Performance Task, the correlation increased to .66 (n=18). Large differences between cross-sectional and longitudinal effect sizes may have resulted from differences in testing conditions affecting motivation, fatigue, or other factors (e.g., 180 minutes of testing for students in the longitudinal study versus 90 minutes of testing for seniors in spring 2006). The correlations were also no doubt suppressed by the less than ideal reliabilities of the effect sizes with both models; therefore, they are quite respectable.

Did most schools have similar freshman-to-senior gain scores, or were there large differences in gain scores among schools?

Figures 1, 2, and 3 show there was noticeable variation in effect sizes across schools. For example, the effect sizes at some schools were several times larger than they were at other schools. Furthermore, the 95 percent confidence interval for a school's effect size (represented by the length of the vertical line through its data point) did not always overlap the effect sizes at other schools, which indicates that some of the differences between schools in effect size were statistically significant.[2] These results are consistent with the thesis that some schools are more effective than others in improving their students' higher order skills.

In addition, the three figures show that the differences in gain scores between schools occur whether gain is measured with a cross-sectional or a longitudinal model. Moreover, the confidence interval for a school's cross-sectional effect size generally corresponded to (overlapped) the interval for its longitudinal effect size. That is, most differences between cross-sectional and longitudinal effect sizes could be explained by unreliability related to sampling variation rather than research design (i.e., they are not necessarily true differences between cross-sectional and longitudinal effect sizes). This pattern was especially clear for the Performance Task. As noted in the previous section, cross-sectional effect sizes for the Analytic Writing Tasks were inflated, and this is reflected in Figure 2. As expected, the Total Score effect sizes tend to fall between the Performance Task and Analytic Writing Task effect sizes.

[2] If 100 random samples are drawn from a population, then about 95 of these samples will have a 95 percent confidence interval that includes the population mean (see Freedman et al., 2007, *Statistics*, 4th edition, p. 385).

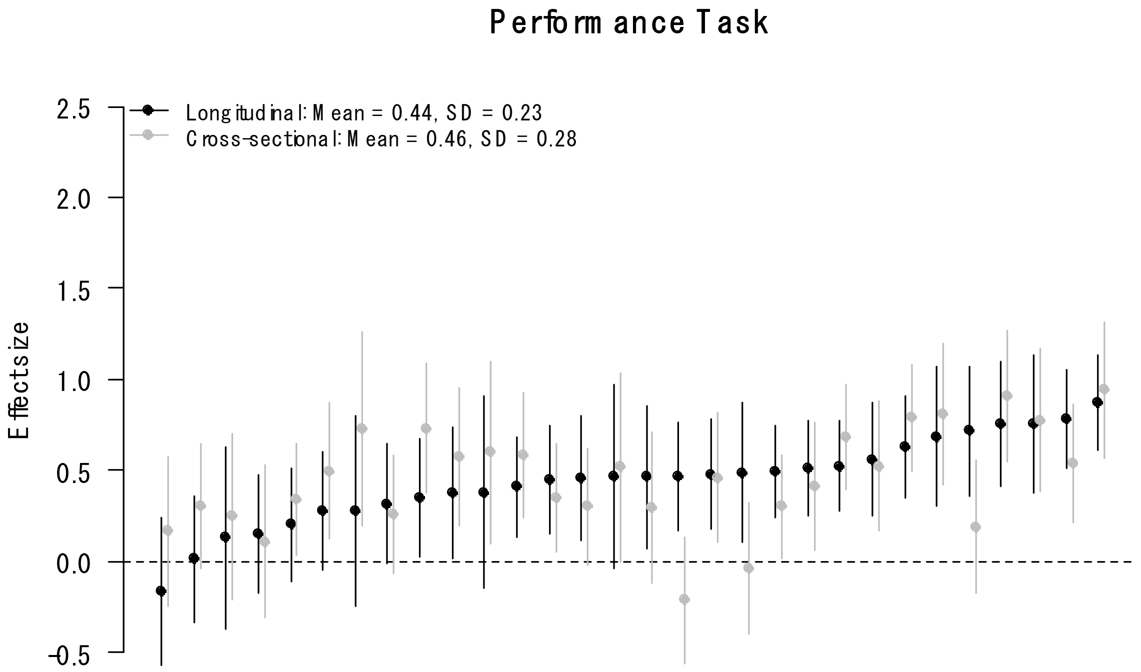

Figure 1. Longitudinal and cross-sectional Performance Task effect sizes (and corresponding 95 percent confidence intervals) for the 30 schools with 25 or more students completing the Performance Task on each testing occasion.

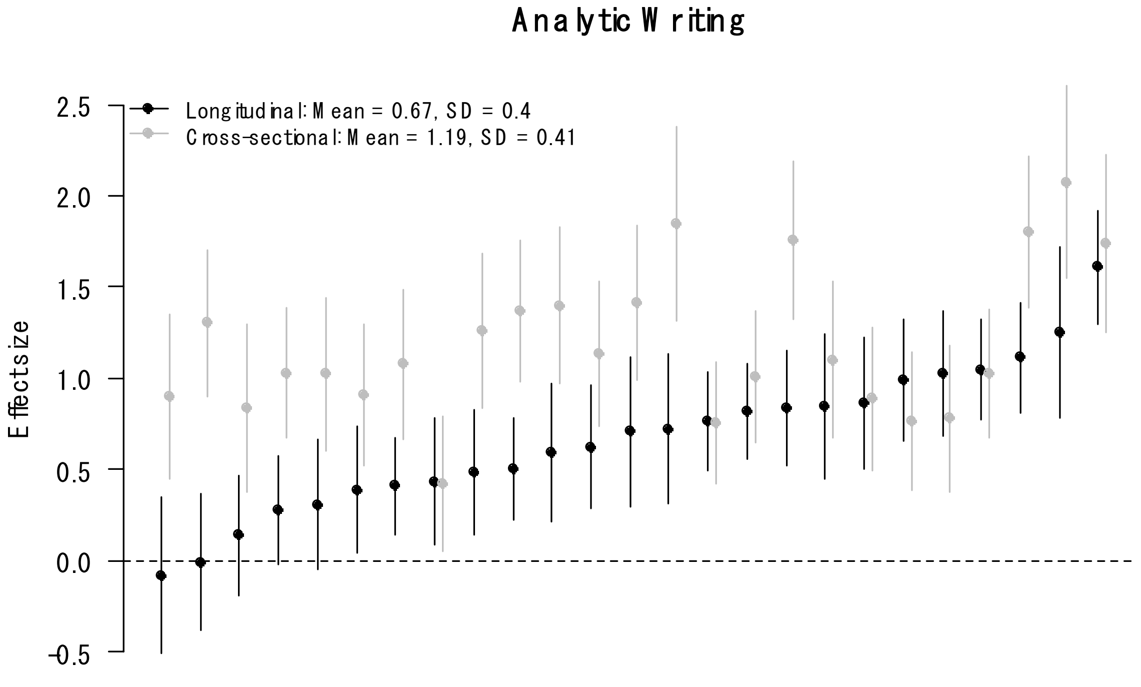

Figure 2. Longitudinal and cross-sectional Analytic Writing effect sizes (and corresponding 95 percent confidence intervals) for the 25 schools with 25 or more students completing the Analytic Writing Task on each testing occasion.

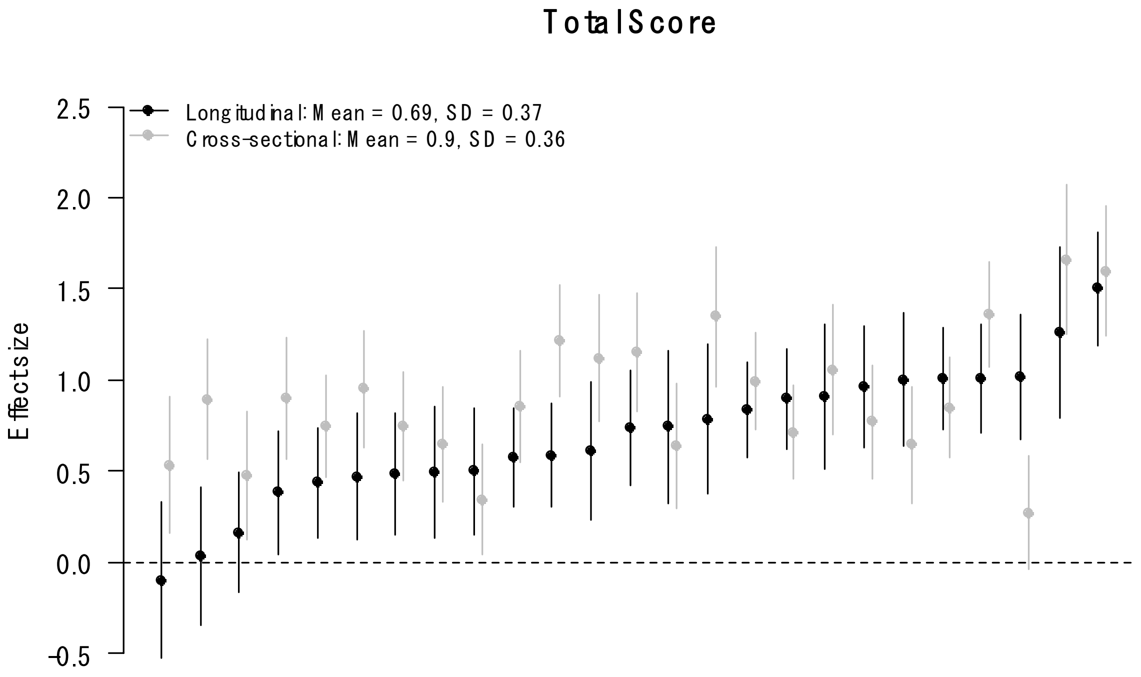

Figure 3. Longitudinal and cross-sectional Total Score effect sizes (and corresponding 95 percent confidence intervals) for the 25 schools with 25 or more students completing the Performance Task and Analytic Writing Task on each testing occasion.

Was the reliability of the cross-sectional CLA gain scores comparable to the reliability of the longitudinal gain scores?

Reliability refers to the consistency of measurement, such as the degree to which students (or schools) tend to receive the same score on a test’s even-numbered questions as they do on its odd-numbered questions. Reliability is generally reported on a 0.00 to 1.00 scale; the higher the coefficient, the greater the reliability.

We investigated school-level reliability using 1,000 random split-sample replications (Klein et al., 2009). Randomly splitting a school’s sample of students into halves is conceptually equivalent to the odd-numbered questions and even-numbered questions on a test at the individual level. The reliability of longitudinal effect sizes on the Performance Task was .55. The cross-sectional reliability was 0.59. Analytic Writing Tasks were omitted from reliability analysis on account of the test fatigue confound mentioned above.

The results of the split-sample reliability analysis suggest that, in this context, cross-sectional and longitudinal designs yield comparable effect size reliability coefficients. There is no clear benefit to score reliability from selecting one design over the other. That said, the reliability of the effect sizes was modest, suggesting that the effect sizes were not as precise an estimate of gain as we would like. For this reason, schools should base assessments of their students' progress on more than one or two CLA administrations worth of data, and they should try to reduce measurement error by increasing sample sizes. These recommendations apply equally to the cross-sectional and longitudinal models. By following these recommendations, schools can ensure that they base consequential decisions on robust and dependable data.

Was the improvement in scores between entering freshmen and "rising juniors" (i.e., students completing their sophomore year) comparable to the improvement between rising juniors and graduating seniors? And, did the answer to this question depend on which type of CLA task was studied?

Figure 4 shows three "box-and-whisker" plots for the Performance Task. The middle line in each plot represents the median longitudinal effect size across schools (i.e., half the schools fall above this line and half below it). Starting at the top of each plot, the other horizontal lines correspond to the maximum, 75th percentile, 25th percentile, and minimum effect sizes (hollow dots indicate outlying points beyond 1.5 "boxlengths"). The plot on the left side of the figure shows the distribution of effect sizes for the first two years of college (between fall 2005 and spring 2007), whereas the middle plot shows the distribution of effect sizes for the last two years (between spring 2007 and spring 2009). The rightmost plot shows the distribution of effect sizes reflecting gain across all four years of college (fall 2005 to spring 2009).

The fact that the leftmost plot is about as high as the middle plot indicates the gains in Performance Task scores during the first two years at Lumina schools were about the same as the gains during the last two years. However, Figures 5 (Make-an-Argument) and 6 (Critique-an-Argument) show that on both Analytic Writing task types, the score gain was much greater during the last two years than it was during the first two years, perhaps because upper-division courses often require more writing. Note that the results presented in Figures 4, 5, and 6 reflect the most restrictive sample (i.e., data from the 20 schools that had at least 25 students completing the entire CLA on all three testing occasions) in order to allow for comparisons across task types.

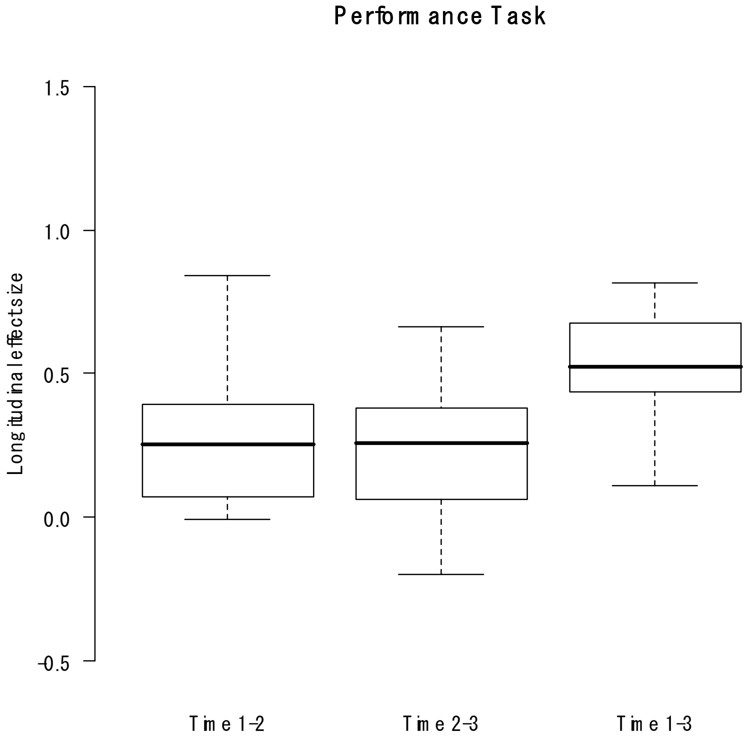

Figure 4. Box-and-whisker plots showing the distribution of longitudinal Performance Task effect sizes across different data collection phases.

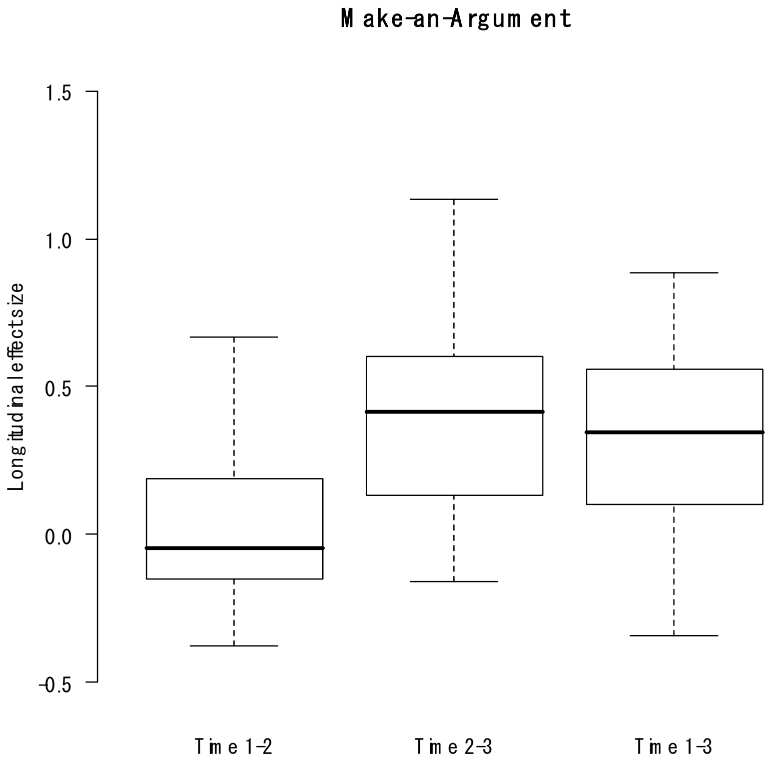

Figure 5. Box-and-whisker plots showing the distribution of longitudinal Make-an-Argument effect sizes across different data collection phases.

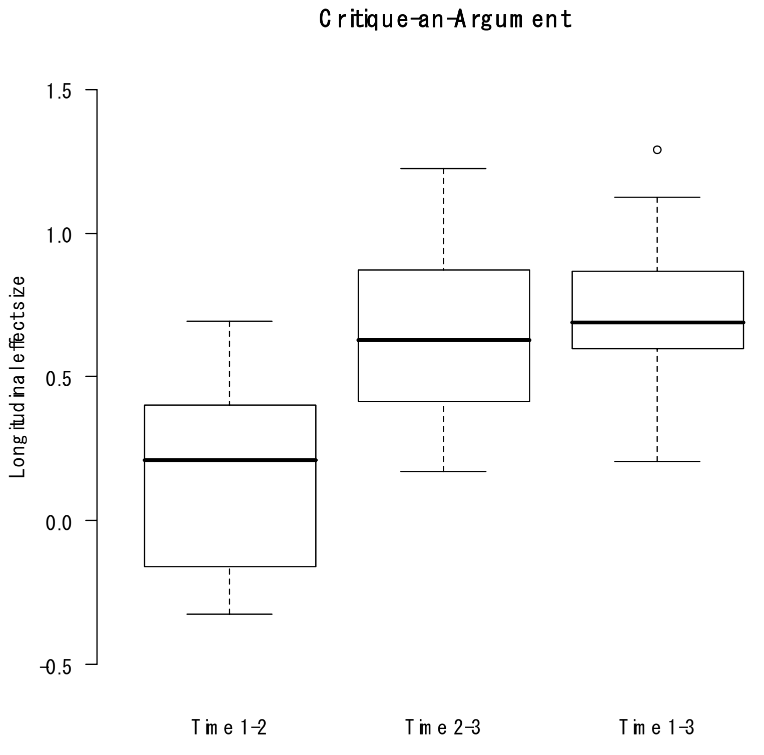

Figure 6. Box-and-whisker plots showing the distribution of longitudinal Critique-an-Argument effect sizes across different data collection phases.

How well do SAT and freshmen CLA scores, by themselves and in combination, predict a student's college GPA?

Unlike the analyses discussed above, our investigation of this question used student-level rather than school-level data. It also involved expanding the CLA database to include each student's cumulative college grade point average (GPA) in their last year in college. To adjust for differences in grading standards across schools, we converted the GPAs within a school to a score distribution that had the same mean and standard deviation as its students' SAT scores. Analyses were conducted with the 1,360 students who had an SAT (or an ACT score converted to the SAT scale) and a GPA at a school with at least 25 students participating in the Lumina study.

The percentage of variance in senior GPAs that can be explained by SAT scores, freshman total CLA score and the combination of SAT and freshman CLA scores was 19, 17, and 23 percent, respectively. Student SAT scores explained only 34 percent of the variance in freshmen CLA scores (i.e., two-thirds of the variance in freshmen CLA scores was *not* explained by SAT scores).[3]

Taken together, these data indicate that a student's SAT score is a slightly (but not statistically significantly) better predictor of that student's cumulative college GPA than is the student's freshman CLA score, but the combination of these two measures is a better predictor of overall GPA than is either one by itself. These data also indicate that there are meaningful differences between what SAT and CLA scores measure.

[3] The correlation between SAT and CLA scores is about 0.90 when the school is the unit of analysis. Equivalently, SAT accounts for 81 percent of the variance in CLA scores when the school is the unit of analysis.

Summary and Conclusions

The Lumina Foundation study reported here administered the CLA to a longitudinal cohort of entering freshmen in the fall of 2005 and then retested these same students in the spring of 2007 (as “rising juniors”) and again in the spring of 2009 when they were graduating seniors. The study also tested a cross-sectional comparison group by administering the CLA in the spring of 2006 to seniors from the same schools as participated in the longitudinal study.

This research found that the cross-sectional differences in Performance Task scores (i.e., between the 2005 freshmen and the 2006 seniors) were consistent with the score differences in the longitudinal cohort of 2005 freshmen that was retested as seniors in 2009. For example, the cross-sectional and longitudinal designs had comparable effect sizes and reliabilities. Moreover, given the limited reliability of differences scores, the correlations between the cross-sectional and longitudinal effect sizes were respectable. Nevertheless, because of the less than stellar reliability of differences scores, we recommend that regardless of which model a school employs to measure change, it should gather at least two or three years worth of data before it uses the results to make educational policy decisions or pat (or whip) itself on the back.

The substantial school and student attrition in the longitudinal cohort is a potentially serious problem. For example, it could be related to unmeasured factors that affect scores, such as student motivation. Longitudinal studies are also inherently more difficult and costly to conduct and their results are often “stale” by the time the data can be analyzed. Thus, given the comparability of the cross-sectional and longitudinal findings, we see no compelling need to adopt the longitudinal model (see also Klein et al., 2008).

One intriguing finding was that while students gained as much on the Performance Task between their first two years in school as during their last two, but almost all the gain on the analytic writing tasks occurred in the last two years. One possible explanation for this finding is that upper-division courses may involve more writing, but there are no doubt other theories. More importantly, the difference in improvement patterns across measures and the differences in effect sizes across schools undercuts the notion that score gains are due to maturation. If maturation were the sole source of growth, we would see similar gains on all tasks and across all institutions.

References

Hardison, C., Hong, E., Chun, M., Kugelmass, H., & Nemeth, A. (2009). *Architecture of the CLA Tasks*. Available online at http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf

Klein, S. (2007). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In "Probability and Statistics: Essays in Honor of David A. Freedman" Deborah Nolan and Terry Speed, editors: IMS Collections, Volume 2, 76-89. Beachwood, OH. Institute for Mathematical Statistics.

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.

Klein, S., Hardison, C. M. et al. (2007). *Can we replace human essay grading with machine grading in large-scale assessment programs?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Klein, S., Liu, O.L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., & Steedle, J. (2009). *Test Validity Study Report*. Prepared under a grant from the Fund for the Improvement of Postsecondary Education. Available online at http://www.voluntarysystem.org/docs/reports/TVSReport_Final.pdf