

CAE

Comparing Longitudinal and Cross-Sectional School Effect Estimates in Postsecondary Education

April 2015

Doris Zahner
Jeffery T. Steedle



Copyright © 2015 Council for Aid to Education

Paper presented at the 2015 National Council on Measurement in Education Annual Meeting, Chicago, IL.

Abstract

Some universities administer standardized tests to estimate school effects on student learning. To that end, many of those institutions choose to gather cross-sectional data (freshmen and seniors in the same academic year) rather than longitudinal data. The effects of this decision on institutional assessment results are generally unknown. The study reported here compared longitudinal and cross-sectional school effect estimates using data from 25 schools that administered the Collegiate Learning Assessment (CLA), an open-ended test of critical-thinking skills. Four different statistical methods were used to estimate school effects, including value-added models commonly used in K-12 and higher education. Correlations between longitudinal and cross-sectional results averaged .47, revealing that data-collection methodology had a large impact on the relative standings of school effect estimates. The effect of using different statistical models was much smaller. Effect sizes, reflecting longitudinal and cross-sectional differences between freshmen and seniors, were similar on average, but differed greatly at a few schools. Effect sizes varied across schools, suggesting that some schools are more successful than others at increasing students' critical-thinking skills.

Keywords: institutional research, value added, critical thinking

Comparing Longitudinal and Cross-Sectional School Effect Estimates in Postsecondary Education

Standardized testing in postsecondary education gained traction as recognition grew for the importance of measuring general academic outcomes (Hart Research Associates, 2009; Kuh & Ikenberry, 2009). Such tests commonly serve dual purposes for a postsecondary institution: gauging academic strengths and weaknesses to inform efforts to improve teaching and learning, and providing evidence of student learning to prospective students, accrediting agencies, policymakers, and stakeholders. Test results are often used to estimate school “effects” on student learning, and these estimates are benchmarked against other participating institutions, sometimes controlling for student or institutional characteristics.

There are different methodologies for collecting learning-outcomes data. Longitudinal data collection is upheld as the gold standard, but longitudinal studies take years to conduct, so results may be stale by the time they are available. Moreover, results may be biased because students' test-taking efforts wane over time, or because a non-representative group of students drops out of the study (Klein, Freedman, Shavelson, & Bolus, 2008). In light of these drawbacks, many postsecondary institutions opt for a cross-sectional approach to gauging learning, which typically involves testing a representative sample of entering freshmen during the fall and a similarly representative sample of graduating seniors the following spring. The Voluntary System of Accountability—a network of public four-year institutions that discloses results from learning outcomes assessments using a shared website template—initially adopted this approach in part because it was “quicker, simpler, and less costly to implement” than longitudinal data collection (Voluntary System of Accountability, 2008, p. 4).

The use of cross-sectional data to measure learning has been contentious, and there are obvious methodological drawbacks to this approach, chief among them being that different groups of students are being compared (Garcia, 2007). To address this limitation, approaches to analyzing cross-sectional data often control for differences in prior ability (Klein, Benjamin, Shavelson, & Bolus, 2007). While some research indicates that controlling for additional student characteristics makes no practical difference for results (e.g., Klein et al., 2008), the freshmen and seniors may still differ on some unaccounted-for variables. Potentially large differences between longitudinal and cross-sectional school effect estimates pose a significant problem for outcomes assessment in higher education, and evidence of their comparability is lacking (Liu, 2011).

To determine whether differences in data-collection methodologies impact school effect estimates, the Lumina Foundation supported the CLA Longitudinal Study, a four-year study using the Collegiate Learning Assessment (CLA), an open-ended test of critical-thinking and

written-communication skills. In this study, 25 postsecondary institutions collected cross-sectional data during the 2005-2006 academic year and longitudinal data between 2005 and 2009. This unique data set provided the opportunity to address the question of whether longitudinal and cross-sectional data yielded similar school effects estimates.

In this study, a variety of analytic approaches, including value-added models applied to K-12 longitudinal data (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004) and postsecondary cross-sectional data (Steedle, 2012), were applied to the available data. Results may inform future decisions about data-collection procedures for institutional assessment programs and supplement the body of research about the development of critical-thinking skills in postsecondary education.

Background

The primary purpose of this study was to examine differences between longitudinal and cross-sectional school effect estimates based on a measure of critical thinking. As background, this section defines critical thinking, describes the CLA as a measure of critical thinking, and summarizes what is currently known about longitudinal and cross-section estimates of growth in critical thinking during college.

Critical Thinking

Skills associated with critical thinking are longstanding desired outcomes of education (Dewey, 1910; Educational Policies Commission, 1961), and in today's society, they are seen as essential for accessing and analyzing the information needed to address the complex, non-routine challenges workers face in the twenty-first century (Hart Research Associates, 2009, 2013; NGA & CCSSO, 2010a, 2010b; The New Commission on the Skills of the American Workforce, 2006; The Secretary's Commission On Achieving Necessary Skills, 1991). In recognition of the central role that critical thinking plays in the twenty-first century, leaders in higher education, business, and government have recommended that such higher-order skills should be assessed at the postsecondary level (Business-Higher Education Forum, 2004; Hart Research Associates, 2006, 2009, 2013; State Higher Education Executive Officers, 2005; U.S. Department of Education, 2006).

Despite variation in definitions of critical thinking, there is general agreement on the core components. The American Philosophical Association's (1990) definition, which reflects the consensus of 200 policy makers, employers, and professors, describes critical thinking as: "Purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference as well as explanation of the evidential, conceptual and methodological considerations on which a judgment is based" (p. 2). Along these lines, Pascarella and Terenzini (2005) operationalize critical thinking as the capacity to "identify central issues and assumptions in an argument, recognize important relationships, make correct references from the data, deduce conclusions from information or data provided, interpret whether conclusions are warranted based on given data, evaluate evidence of authority, make self-corrections, and solve problems" (p. 156).

Although postsecondary students most often employ critical-thinking skills in the context of their chosen fields of study, intellectual skills that transcend academic majors are commonly upheld as important outcomes of higher education (Pascarella & Terenzini, 2005). In accord with this idea, 78% of the Association of American Colleges and Universities member institutions report having a "common set of intended learning outcomes for all their undergraduate students," and "the skills most widely addressed are writing, critical thinking, quantitative reasoning, and oral communication skills" (Hart Research Associates, 2009, p. 2).

Collegiate Learning Assessment

The CLA is a standardized testing program utilized by postsecondary schools to gauge student's critical-thinking skills in accordance with consensus definitions of critical thinking. Since the CLA was first administered on a large scale in 2004, approximately 400,000 students at 700 colleges and universities in the United States and internationally (AHELO, 2012a, 2012b) have participated. In fall 2013, CAE launched CLA+, which is an enhanced version of the CLA that is valid and reliable at the individual student level (Zahner, 2013). The majority of participating CLA and CLA+ institutions opt to collect cross-sectional data from entering freshmen and graduating seniors during the same academic year. Student scores are aggregated, and the

relationship between freshman and senior average performance provides a school effect estimate that can be benchmarked against other schools participating in the CLA. Prior investigations revealed that aggregate CLA scores are reliable and correlate with other measures of critical thinking (Klein et al., 2007; Klein et al., 2009; Steedle, 2012). Additional details about the format of the CLA are provided in the method section below.

Research on Gains in Critical Thinking

The thesis that attending college improves critical thinking was supported by a meta-analysis of 27 studies, many of which employed the Watson-Glaser Critical Thinking Appraisal (McMillan, 1987). Additional evidence was provided by Pascarella and Terenzini's (1991) review of the effects of higher education in which they estimated the average difference in critical-thinking skills between freshmen and seniors as approximately 1.0 standard deviation. In research from the 1990s, Pascarella and Terenzini (2005) found that most cross-sectional research on critical thinking revealed gains smaller than 1.0 standard deviation, with mean differences between freshmen and seniors falling between 0.55 and 0.65 standard deviations. Longitudinal results were sparse, but they tended to suggest even smaller gains in critical thinking (e.g., 0.25, 0.15 and 0.37 over three years of college).

In another study, a longitudinal increase of 0.84 standard deviations from freshman to senior year was observed for students taking ACT's College Outcomes Measures Program (COMP) (Pike, 1992). Ohio University conducted 11 consecutive longitudinal studies using COMP between 1981 and 1995 and obtained freshman-senior effect sizes ranging from 0.51 to 1.15 with an average of 0.73 (Ohio University Office of Institutional Research, 1996). In the National Study of Student Learning, which administered ACT's Collegiate Assessment of Academic Proficiency (CAAP), the CAAP critical-thinking scores of full-time students increased by 0.41 standard deviations more than they increased for part-time students (Pascarella, Bohr, Nora, & Terenzini, 1996).

Recent evidence comes from two large-scale longitudinal studies. Arum and Roksa (2011) examined data from the CLA Longitudinal Study, which included more than 3,000 students at 29 institutions. They reported an average increase of 0.18 standard deviations on the CLA during the first two years of college and a four-year effect size of 0.47 (Arum, Roksa, & Cho, 2011). In the Wabash National Study, CAAP was administered to samples of students at 17 four-year institutions on three occasions: upon entering college, at the end of the first year, and at the end of the fourth year (Pascarella, Blaich, Martin, & Hanson, 2011). The average gain on CAAP critical thinking was 0.11 standard deviations during the first year and 0.44 standard deviations over four years.

These cited studies were all conducted with an assortment of tests at a variety of institutions, so they do not provide a good evaluation of the similarity between longitudinal and cross-sectional estimates of school effects. Even if such estimates were similar on average across studies, what matters most to institutional researchers is whether longitudinal and cross-sectional estimates are similar for individual institutions. The research reported here directly addressed this issue by examining the correspondence of longitudinal and cross-sectional school effect estimates. Specifically, the results of this study address the following questions: (1) How are the relative standings of school effect estimates affected by differences in statistical estimation and data-collection methods? And, (2) are longitudinal and cross-sectional effect sizes comparable?

Before proceeding, two important caveats about this study should be addressed. First, the results of this study cannot indicate which data collection method (longitudinal or cross-sectional) provides a more accurate measure of student learning because there exists no objective criterion against which results can be compared (e.g., the average gain of a representative sample of students who attended an institution from freshman to senior year). It is commonly assumed that longitudinal and cross-sectional results should be similar, and this investigation was focused on testing that assumption. Second, the data analyzed in this study were authentic, but they were not optimal. That is, it would be ideal if sample sizes were very large and if attrition was negligible. These conditions were not met, but these data were used by actual postsecondary institutions to estimate student learning. Results, therefore, provide a realistic depiction of differences in school effects estimates that can be expected from different data collection methods.

Method

Measure

CLA testing took place in computer labs under the supervision of trained proctors. Students first completed a 90-minute performance task in which they analyzed a document library containing a mixture of trustworthy and unreliable information related to a real-world scenario. Students then demonstrated their problem-solving skills by proposing a solution and supporting it, arguing against alternative solutions, and proposing additional research. Next, students completed a 75-minute analytic writing task, which required them to construct an argument on a provided topic and write a critique of an argument written by someone else. Sample CLA tasks are available at www.cae.org/cla.

Responses were evaluated using a combination of trained scorers and automated essay scoring engines. Each response was scored using rating scales that captured aspects of critical thinking and analytic writing. Correlations between human scorers were typically around .85, and correlations between human scorers and automated essay scoring engines approached .90 (Steedle & Elliot, 2012). Scores on different tasks were converted to a common scale to adjust for differences in task difficulty. A total score equal to the average of the performance task and the analytic writing task was analyzed in this study.

Test Administration

Students who participated in the CLA Longitudinal Study tested as freshmen in fall 2005 and again as seniors in spring 2009. Participating schools also tested a sample of their spring 2006 graduating seniors, which made it possible to estimate cross-sectional differences between freshmen and seniors in the same academic year. Note that the spring 2006 seniors took either a performance task or an analytic writing task. The longitudinal cohort was atypical in that they completed both sections of the CLA.

Subjects

To help ensure the reliability of aggregate test scores, schools were only included if they had 30 or more participating students in each class. Of the 25 schools with sufficient sample sizes, 13 were research universities, five were master's colleges and universities, and seven were baccalaureate colleges. Ten schools had a total enrolment under 10,000 students, 12 were public institutions, and two were classified as Historically Black Colleges and Universities. They had admissions rates ranging from 26% to 95% (median 67%) and six-year graduation rates ranging from 41% to 93% (median 67%).

Analyses only included students with CLA scores and college admissions test scores—that is, SAT or ACT scores converted to the SAT scale (ACT, 2008). Three groups of students were examined:

1. 1,669 students who tested as freshmen in fall 2005 and as seniors in spring 2009 (the longitudinal cohort)
2. 5,114 students who tested as freshmen in fall 2005 as part of the CLA Longitudinal Study (the freshman cross-sectional comparison group, which included the 1,669 students in the longitudinal cohort)
3. 1,508 students who tested as seniors in spring 2006 (the senior cross-sectional comparison group)

The longitudinal cohort was 65% female and 71% White, with 89% reporting that English is the primary language spoken at home (Table 1). The schools in this study experienced student attrition rates of between 27% and 86% (median 67%) from the first phase (fall 2005) of the longitudinal study. Compared to students who began but did not complete the longitudinal study, students who persisted through all three phases of the longitudinal study tended to have higher admissions test scores, and they were more likely to be White and speak English at home. Consequently, conclusions about longitudinal gains on the CLA should be restricted to students like those in the longitudinal cohort. Similar concerns might be raised about differences between the freshman and senior cross-sectional comparison groups, but the two groups in the cross-sectional cohort were much more similar in terms of average prior ability and demographics.

Table 1 Sample Demographics						
Sample	N	Mean SAT	Female	White	English Spoken at Home	Mean Freshman CLA
Cross-sectional freshmen (fall 2005)	5114	1148	64%	67%	87%	1128
Longitudinal cohort (fall 2005-spring 2009)	1669	1179	65%	71%	89%	1151
Cross-sectional seniors (spring 2006)	1508	1151	65%	66%	87%	--
Longitudinal cohort attrition (fall 2005)	3445	1133	64%	65%	87%	1117

Models for Estimating School Effects

Four methods for estimating school effects were employed in this study. Some methods are used in operational and research contexts as value-added models, which are upheld as fair methods of estimating the effects of schools or teachers because they control for student or school characteristics (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). For the sake of comparability between methods, only adjustments for students' prior achievement levels, as measured by their SAT or converted ACT scores, were made in this study. Other research has examined the effects of adjusting for additional variables, and such effects tend to be very small as long as adjustments are also made for prior achievement (Ballou, Sanders, & Wright, 2004; Hægeland & Kirkebøen, 2008; Klein et al., 2008; Tekwe et al., 2004). Thus, adjusting for additional variables would not be expected to impact results anyway.

Methods for cross-sectional data. A *cross-sectional effect size* indicates the mean difference in CLA performance between different groups of freshmen and seniors expressed in standard deviation units. The cross-sectional effect size at school j is calculated using the following equation.

$$d_{cross,j} = \frac{\bar{X}_{S,j} - \bar{X}_{F,j}}{s_{X,j}}$$

where $\bar{X}_{S,j}$ and $\bar{X}_{F,j}$ are the respective mean CLA scores of seniors and freshmen at school j (within the same academic year), and $s_{X,j}$ is the pooled sample standard deviation of CLA scores at school j . A notable limitation of the cross-sectional effect size is that it may be biased due to differences in academic ability between the tested freshmen and seniors.

The CLA *value-added model* (Steedle, 2012) is a hierarchical linear model that treats senior students (level 1) as nested within schools (level 2). At level 1, student i at school j 's SAT score (Y_{ij}) is employed as a covariate for his or her CLA score (X_{ij}).

$$\text{Level 1: } X_{ij} = \beta_{0j} + \beta_{1j}(Y_{ij} - \bar{Y}_{S,j}) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{Y}_{S,j} + \gamma_{02}\bar{X}_{F,j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

The SAT scores in level 1 are group-mean centred, which makes β_{0j} equal to the senior mean CLA score at school j . At level 2, the senior mean CLA score (β_{0j}) is modelled as a function of the mean SAT score of participating seniors ($\bar{Y}_{S,j}$) and the mean CLA score of participating freshmen ($\bar{X}_{F,j}$). The level-2 residual is the value-added score for school j .

$$VA_{CLA,curr,j} = u_{0j}$$

This score indicates whether the participating seniors at school j performed above or below expectations based on their mean SAT score and the mean CLA score of the participating freshmen.

Methods for longitudinal data. A *longitudinal effect size* indicates the mean difference in CLA performance between freshman and senior years for a single group of students expressed in standard deviation units. Formula (1) applied to longitudinal data provides a longitudinal effect size.

$$d_{long,j} = \frac{\bar{X}_{S,j} - \bar{X}_{F,j}}{s_{X,j}}$$

The *random-effects model* (Meyer, 1996; Raudenbush & Willms, 1995) employs a hierarchical linear model with students nested in schools. A perceived advantage of the random-effects model is that it provides empirical Bayes estimates of school effects (Raudenbush & Bryk, 2002). Such estimates are “shrunk” toward the mean according to the precision of an individual school’s effect and the overall variance of the effects.

At level 1, student i at school j ’s SAT score (Y_{ij}) and prior CLA score (X_{ij}^*) are employed as covariates for his or her senior CLA score (X_{ij}). Both variables are grand-mean centred, which makes β_{0j} equal to the senior mean CLA score adjusted for SAT and prior CLA scores.

$$\text{Level 1: } X_{ij} = \beta_{0j} + \beta_{1j}(Y_{ij} - \bar{Y}_S) + \beta_{2j}(X_{ij}^* - \bar{X}^*) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

The shrunk level-2 residual is the value-added score for school j .

$$VA_{RE,j} = \lambda_j u_{0j}$$

In this formula, u_{0j} indicates how a school’s adjusted senior mean CLA score deviates from the grand mean adjusted CLA score. The λ_j coefficient is a shrinkage weight factor that is closer to 1.0 when schools have large sample sizes or little within-school variance.

Other value-added models for longitudinal data such as the *mean model* (e.g., Ladd & Walsh, 2002) and the *fixed-effects model* (e.g., Coleman, Campbell, & Kilgore, 1982) were considered in this study, but their correlations with the random-effects model approached 1.00, which is consistent with prior research (Jakubowski, 2008; Ladd & Walsh, 2002). Therefore, any conclusions related to the random-effects model would also be applicable to other models.

Analysis

The analysis was designed to first address the question “How are the relative standings of school effect estimates affected by differences in statistical estimation and data-collection methods?” Note that the focus on relative standings (rather than absolute levels) was necessitated by the fact that different school effect estimates are generally on different scales of measurement. Correlations were computed between results from all the estimation methods described above. Low correlations would suggest that choices between estimation methods and data-collection methods have large effects on the relative standings of school effect estimates. However, an additional question would still remain: “Were the correlations lower because of differences in estimation methods or because of differences in data-collection methods?”

To address this question, the CLA model and the random-effects models were applied to the “wrong” data sets. That is, the CLA model was applied to longitudinal data, and the random-effects model was applied to cross-sectional data. Note that prior CLA scores could not be included in the random-effects model, leaving SAT as the only level-1 covariate. Correlations were, again, computed to see which variable (the estimation method or the data-collection method) affected correlations the most.

The subsequent analysis addressed the question “Are longitudinal and cross-sectional effect

sizes comparable?” These analyses focused on effect sizes because most prior research has reported effects sizes and, unlike the other school effect estimates, the effect sizes were on a consistent scale of measurement. To address the first question, longitudinal and cross-sectional effect sizes were compared on average and also within individual schools. Standard errors of the effect sizes were computed to illustrate expected variability due to sampling error (Hedges & Olkin, 1985).

Results

Relative Standings of School Effect Estimates

The first set of results includes correlations indicating the degree of consistency in the relative standings of the school effect estimates from the four estimation methods (Table 2). The correlation between school effect estimates based on cross-sectional data was .84 ($p < .001$). The longitudinal effect size correlated .89 ($p < .001$) with the random-effects value-added model.

	ES.CS	CLA.CS	ES.Long	RE.Long
Cross-sectional effect size (ES.CS)	1.00			
CLA value-added model (CLA.CS)	.84***	1.00		
Longitudinal effect size (ES.Long)	.53**	.46*	1.00	
Random-effects model (RE.Long)	.39	.51**	.89***	1.00

*** $p < .001$, ** $p < .01$, * $p < .05$

The italicized values in Table 2 address the main focus of this study: comparing longitudinal and cross-sectional school effect estimates. The longitudinal effect size correlated .53 ($p < .01$) with the cross-sectional effect size and .46 with the CLA model ($p < .05$). The random-effects model correlated significantly with the CLA model ($r = .51$, $p < .01$), but not with the cross-sectional effect size ($r = .39$). In all cases, these cross-methodological correlations were lower than the within methodology correlations.

Relative Effects of Estimation Methods and Data Collection Methods

The results presented in this section illustrate the relative effects of the choices between estimation methods and data-collection methods on the correlation between school effect estimates. The two lowest values in Table 3 (.51 and .52) reflect correlations between different models applied to different data sets. All other correlations in Table 3 are higher because they reflect shared variance on account of using the same model or the same data set.

	CLA.CS	CLA.Long	RE.CS	RE.Long
CLA value-added model (CLA.CS)	1.00			
CLA model with longitudinal data (CLA.Long)	.56**	1.00		
Random-effects model with cross-sectional data (RE.CS)	.90***	.52**	1.00	
Random-effects model (RE.Long)	.51**	.88***	.65***	1.00

*** $p < .001$, ** $p < .01$, * $p < .05$ First consider the same model applied to different data sets.

The CLA model, applied separately to longitudinal and cross-sectional data sets, correlated .56 ($p < .01$) with itself. The random-effects model applied to different data sets correlated .65 ($p < .001$) with itself. The modest differences between these correlations and the lowest correlations (.51 and .52) indicate the positive effect on the correlation from using the same statistical model.

Now consider different models applied to the same data set. Applied only to cross-sectional data, the CLA model and the random-effects model correlated .90 ($p < .001$). When applied to

longitudinal data, the two models correlated .88 ($p < .001$). These correlations are much higher than other correlations in Table 3, suggesting that lower correlations were caused more by the choice of data than the choice of statistical model.

Relative Magnitudes of Longitudinal and Cross-sectional Effect Sizes

Results presented here illustrated correspondence in magnitude of the longitudinal and cross-sectional effect sizes. Across the 25 schools, the mean longitudinal and cross-sectional effect sizes were 0.67 and 0.66, respectively. Standard errors for the effect sizes were used to compute precision-weighted mean effect sizes, which take into account varying sample sizes across schools (Hedges & Olkin, 1985). The meta-analysis effect size was 0.71 for the longitudinal data and 0.63 for the cross-sectional data.

However, the similarity in mean longitudinal and cross-sectional effect sizes hides larger differences for individual schools. The moderate correlation between longitudinal and cross-sectional effect sizes ($r = .53$; $p < .01$) is apparent in Figure 1, which shows 95% confidence intervals for the effect sizes. In many of the schools, differences between longitudinal and cross-sectional effect sizes were within the range of expected variability due to sampling error, but there were several schools with large differences between their longitudinal and cross-sectional effect sizes. Indeed, excluding the schools with the two largest differences, the correlation between longitudinal and cross-sectional effect sizes was .74 ($p < .001$), reflecting a doubling of the variance accounted for (i.e., an increase of R^2 from .27 to .55).

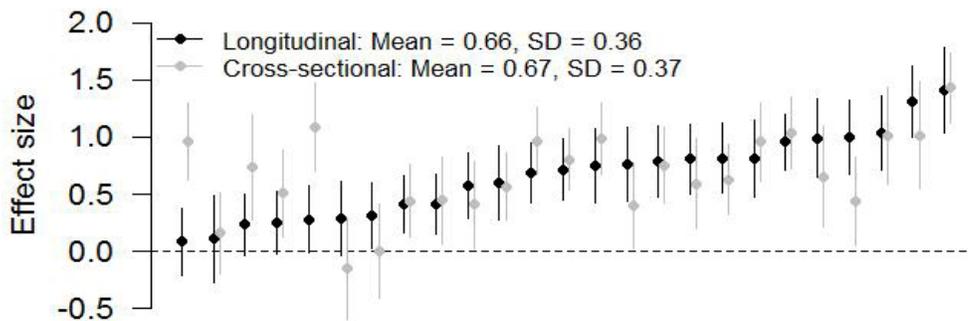


Figure 1. Longitudinal and Cross-Sectional Effect Sizes with 95% Confidence Intervals.

When comparing the rankings of schools by longitudinal and cross-sectional effect sizes, 15 schools (60%) differed by five or fewer rank positions, five schools (20%) differed by six to 10 rank positions, and the remaining five schools differed by 11 or more rank positions. The largest change was 17 positions, reflecting a longitudinal effect size of 0.97 and a cross-sectional effect size of 0.08. Considering what was known about the demographics of participating institutions, there did not appear to be any distinguishing characteristic of institutions with large differences between longitudinal and cross-sectional effect sizes.

Another possible source of the differences between longitudinal and cross-sectional effect sizes was the pooled sample standard deviations. That is, the denominator in the effect-size calculations could have differed between the longitudinal and cross-sectional data sets. Considering that the freshman cross-sectional comparison group was a larger and, likely, more diverse sample than the longitudinal cohort or the cross-sectional seniors, one might expect the pooled sample standard deviation to be greater in the cross-sectional comparison. In fact, the average denominator in the cross-sectional effect sizes was 15 points higher than it was for the longitudinal effect sizes. This difference amounted to approximately 0.08 standard deviations. Thus, observed differences between longitudinal and cross-sectional effect sizes would be somewhat different if the samples were more similar.

Discussion

Summary

The study reported here included an examination of CLA results from a longitudinal cohort and cross-sectional comparison groups. Four statistical methods for estimating school effects were applied to the data. Results from the longitudinal models correlated well with each other, as did results from the cross-sectional models. However, correlations were much lower between longitudinal and cross-sectional results. Compared to the choice of statistical model, the choice of data-collection methodology had a greater impact on the relative standings of school effects estimates.

On average, the longitudinal and cross-sectional effect sizes were quite similar, and they were also comparable to results from previous research (Pascarella & Terenzini, 2005). Observed differences between longitudinal and cross-sectional effect sizes were no greater (and often smaller) than those that have been observed in consecutive longitudinal studies (e.g., Ohio University Office of Institutional Research, 1996, where consecutive study differences ranged from .02 to .41). However, in some schools, differences between longitudinal and cross-sectional effect sizes far exceeded the range of expected variability in results due to sampling error.

Challenges and Recommendations

Issues related to sample representativeness call into question the generalizability of longitudinal school effects estimates. In this study, students in the longitudinal cohort tended to have higher SAT scores and were more likely to be White and speak English at home compared to students who began but did not complete the longitudinal study. It is not known why these students dropped out of the study or whether they would have gained as much on the CLA during college as the longitudinal cohort did. Recruiting a longitudinal cohort of students who are likely to graduate would reduce the problem of attrition, but not necessarily the problem of sample representativeness.

Cross-sectional data collection suffers from a different shortcoming related to sample representativeness: the freshmen and seniors are drawn from different populations, and uncontrolled-for differences between the samples could bias comparisons. To minimize this concern, one might consider targeting a sample of seniors that is similar to the freshman class in terms of prior ability and demographics. Alternatively, one could draw a sample of freshmen similar to the senior class, and this could possibly produce results more similar to those from longitudinal data collection, because results would be based on students like those who persist to graduation.

As noted previously, results from this study cannot indicate which data-collection procedure was more accurate, because the criterion results were unknowable. From this study, one can only conclude that longitudinal and cross-sectional results sometimes differ. Thus, schools must carefully consider the trade-offs inherent in the decision between longitudinal and cross-sectional data collection and unique contextual factors that could influence the validity of interpreting results. For example, the interpretation of cross-sectional results requires the assumption that little has changed at an institution over the past three years. However, recent changes that affect only freshmen (e.g., new admissions policies) would violate this assumption and introduce a source of possible bias in cross-sectional results.

Other possible contextual factors include inconsistent sampling and test administration procedures and systematic differences in motivation between samples. Indeed, context effects related to student recruiting and test administration procedures have been hypothesized as the cause for inconsistent CLA results across years (Hosch, 2010). Unlikely results in this study, such as negative effect sizes, illustrate the need to interpret results in light of these factors.

It is recommended that, regardless of which approach is employed to estimate school effects, multiple measures of student learning be administered to see how results converge and diverge. Data from administrations in consecutive years should also improve confidence in consequential decisions about educational programs.

“2+2” Longitudinal Study

Neither longitudinal nor cross-sectional data collection is ideal because of various pragmatic challenges and threats to measurement quality. To address some of these issues, a hybrid approach might be considered. In the fall of a given year, an institution could test a sample of entering freshmen and a sample of juniors beginning their third year of study. Then, approximately 18 months later, the institution could retest the same students as end-of-year sophomores and graduating seniors. The separate longitudinal cohorts could be used to estimate gains in the first two years of college, the last two years of college, and, by adding them, an estimate of four-year longitudinal gains.

This hybrid approach has several notable advantages. It is based wholly on longitudinal data, and would not therefore suffer from the major weakness of cross-sectional data collections (i.e., uncontrolled-for differences between tested samples). From a pragmatic standpoint, institutions would only have to wait two years to get results. Moreover, this design would only entail two test administrations over the course of two academic years, and should cost approximately the same as two consecutive years of cross-sectional data collection. Sample representativeness may be better with this design for two reasons. First, because the study is only two years long, attrition should be less problematic. Second, this design can accommodate transfer students into the junior-senior year cohort. Such students cannot possibly be included in four-year longitudinal studies and should not be included in cross-sectional studies.

Directions for Future Research

Large differences between longitudinal and cross-sectional results were occasionally observed in this study. Future research might gather information about sample recruitment, test administration procedures, and student self-reported effort in order to possibly explain unexpected results like negative effect sizes or large differences in results across different studies.

In addition to collecting information about test administration conditions, future research should focus on correlates of student learning and examine possible curricular interventions to improve the impact of college on general, higher-order outcomes. Along these lines, other analyses conducted with data from the CLA Longitudinal Study revealed a positive association between high expectations for reading and writing in college courses and improvement on the CLA (Arum & Roksa, 2011). Several schools in this study had effect sizes approaching or exceeding 1.0, and numerous schools in CLA archival data show consistently large, positive cross-sectional effect sizes. These schools (along with those showing consistently low effect sizes) provide a ripe opportunity for studying institutional policies and practices that foster student gains in critical-thinking skills.

References

- ACT. (2008). ACT-SAT Concordance. Retrieved July 21, 2011, from <http://www.act.org/aap/concordance/>
- AHELO. (2012a). AHELO feasibility study interim report. Paris: OECD.
- AHELO. (2012b). Testing students and university performance globally: OECD's AHELO. Retrieved February 28, 2014, 2014, from <http://www.oecd.org/edu/skills-beyond-school/testingstudentanduniversityperformancegloballyoecdshahelo.htm>
- American Philosophical Association. (1990). Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction - "The Delphi Report". Committee on Pre-College Philosophy. Millbrae, CA: The California Academic Press.
- Arum, R., & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.
- Arum, R., Roksa, J., & Cho, E. (2011). Improving Undergraduate Learning: Findings and Policy Recommendations from the SSRC-CLA Longitudinal Project. Brooklyn, NY: Social Science Research Council.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-65.
- Business-Higher Education Forum. (2004). Public Accountability for Student Learning in Higher Education: Issues and Options. Washington, DC: American Council on Education.
- Coleman, J. S., Campbell, T. E., & Kilgore, S. B. (1982). High Achievement: Public, Catholic, and other private schools compared. New York, NY: Basic.
- Dewey, J. (1910). *How We Think*. Boston, MA: D.C. Heath.
- Educational Policies Commission. (1961). *The Central Purpose of American Education*. Washington, DC: National Education Association.
- Garcia, P. (2007). *How to Assess Expected Value Added: The CLA Method*. Paper presented at the California Association of Institutional Research Conference, Monterey, CA.
- Hægeland, T., & Kirkebøen, L. (2008). School Performance and Value-Added Indicators - What is the Importance of Controlling for Socioeconomic Background?: *Background Report for the OECD Project on the Development of Value-added Models in Education Systems*.
- Hart Research Associates. (2006). How Should Colleges Prepare Students to Succeed in Today's Global Economy? - Based on Surveys Among Employers and Recent College Graduates. Washington, DC: Hart Research Associates.
- Hart Research Associates. (2009). Learning and Assessment: Trends in Undergraduate Education - A Survey Among Members of The Association of American Colleges and Universities. Washington, DC: Hart Research Associates.
- Hart Research Associates. (2013). It takes more than a major: Employer priorities for college learning and student success. Washington, DC: AAC & U.
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press.
- Hosch, B. J. (2010). *Time on Test, Student Motivation, and Performance on the Collegiate Learning Assessment: Implications for Institutional Accountability*. Paper presented at the Association for Institutional Research Annual Forum, Chicago, IL.
- Jakubowski, M. (2008). Implementing Value-Added Models of School Assessment *EUI Working Papers RSCAS 2008/06*. Florence, Italy: European University Institute, Robert Schuman Centre for Advanced Studies.
- 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., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., . . . Steedle, J. (2009). Test Validity Study (TVS) Report. Supported by the Fund for the Improvement of Postsecondary Education. from http://www.cae.org/content/pdf/TVS_Report.pdf
- Kuh, G., & Ikenberry, S. (2009). More Than You Think, Less Than We Need: Learning Outcomes Assessment in American Higher Education. Champaign, IL: National Institute for Learning Outcomes Assessment.
- Ladd, H. F., & Walsh, R. P. (2002). Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right. *Economics of Education Review*, 21(1), 1-17.
- Liu, O. L. (2011). Outcomes Assessment in Higher Education: Challenges and Future Research in the Context of Voluntary System of Accountability. *Educational Measurement: Issues and Practice*, 30(3), 2-9.

- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating Value-Added Models for Teacher Accountability. Santa Monica, CA: RAND Corporation.
- McMillan, J. H. (1987). Enhancing College Students' Critical Thinking: A Review of Studies. *Research in Higher Education*, 26(1), 3-29.
- Meyer, R. H. (1996). Value-Added Indicators of School Performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's Schools: The Role of Incentives* (pp. 197-223). Washington, DC: National Academies Press.
- NGA, & CCSSO. (2010a). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers,.
- NGA, & CCSSO. (2010b). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers,.
- Ohio University Office of Institutional Research. (1996). General Education Outcomes: The College Outcomes Measures Program (COMP) at Ohio University 1981-1985. Athens, OH: Ohio University.
- Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How Robust Are the Findings of Academically Adrift? *Change*, 43(3), 20-24.
- Pascarella, E. T., Bohr, L., Nora, A., & Terenzini, P. T. (1996). Is Differential Exposure to College Linked to the Development of Critical Thinking? *Research in Higher Education*, 37(2), 159-174.
- Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco, CA: Jossey-Bass.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research*. San Francisco, CA: Jossey-Bass.
- Pike, G. R. (1992). *Lies, Damn Lies, and Statistics Revisited: A Comparison of Three Methods of Representing Change*. Paper presented at the Annual Forum of the Association for Institutional Research, San Francisco, CA.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications, Inc.
- Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307-335.
- State Higher Education Executive Officers. (2005). Accountability for Better Results: A National Imperative for Higher Education. Boulder, CO: State Higher Education Executive Officers.
- Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education*, 37(6), 637-652.
- Steedle, J. T., & Elliot, S. (2012). The Efficacy of Automated Essay Scoring for Evaluating Student Responses to Complex Critical Thinking Performance Tasks. New York, NY: Council for Aid to Education.
- Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., . . . Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11-36.
- The New Commission on the Skills of the American Workforce. (2006). Tough Choices or Tough Times. Washington, DC: National Center on Education and the Economy.
- The Secretary's Commission On Achieving Necessary Skills. (1991). What Work Requires of Schools: A SCANS Report for America 2000. Washington, DC: U.S. Department of Labor.
- U.S. Department of Education. (2006). A Test of Leadership: Charting the Future of US Higher Education. Washington, DC: U.S. Department of Education.
- Voluntary System of Accountability. (2008). Information on learning outcomes measures. <http://www.voluntarysystem.org/docs/cp/LearningOutcomesInfo.pdf>
- Zahner, D. (2013). Reliability and validity of the CLA+. New York: CAE.