High Educ (2011) 61:445–461
DOI 10.1007/s10734-010-9340-8

Value-added assessment in higher education: a comparison of two methods

Ou Lydia Liu

Published online: 8 May 2010

Abstract Evaluation of the effectiveness of higher education has received unprecedented attention from stakeholders at many levels. The Voluntary System of Accountability (VSA) is one of the initiatives to evaluate institutional core educational outcomes (e.g., critical thinking, written communication) using standardized tests. As promising as the VSA method is for calculating a valueadded score and allowing results to be comparable across institutions, it has a few potential methodological limitations. This study proposed an alternative way of value-added computation which takes advantage of multilevel models and considers important institution-level variables. The institutional value-added ranking was significantly different for some of the institutions (i.e., from being ranked at the bottom to performing better than 50% of the institutions) between these two methods, which may lead to substantially different consequences for those institutions, should the results be considered for accountability purposes.

Keywords ETS Proficiency Profile · Outcomes assessment · Value-added · Voluntary system of accountability

Introduction

In a global economy, a good college education is not only a springboard to opportunity, but also a prerequisite for our young generation to survive and thrive in the twenty-first century. The quality of a nation's higher education system significantly contributes to its international competitiveness. Many countries have made developing higher education one of their top priorities. For example, with the Chinese government's heavy investment in higher education, the number of college graduates increased from 829,070 to 1,594,130 from 1997 to 2007 in China, almost doubled in the last 10 years (www.moe.gov.cn). The importance of higher education is also being increasingly recognized in the United States. The US higher education community is striving to meet a new national goal that by year

O. L. Liu (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: lliu@ets.org

2020, the United States should have the highest proportion of college graduates in the world.

As massive resources are being invested into expanding American higher education, crucial questions arise as to what determines the quality of higher education and what should be done to facilitate reforms and improve efficiency at institutions. Accountability in higher education has become a key area of interest following the attention accountability received in K-12 education. To improve the current state of US higher education, a Commission on the Future of Higher Education was established by the Department of Education under the leadership of former Secretary of Education Spellings. This commission identified four areas that require urgent reform in higher education: expanding access to postsecondary education, increasing affordability for students and families, improving the quality of college instruction, and achieving improved accountability (US Department of Education, National Center for Education Statistics 2006). The Commission points out that there is a dire need to develop accountability mechanisms to ensure that quality education is made available to students. It also emphasizes the importance of mechanisms that could produce comparable results across institutions. The Commission calls for outcomes measures that can provide value-added information about student progress in college while taking into account student baseline achievement.

Over the past decade, attention to outcomes assessment for accountability purposes in higher education has increased exponentially. This attention comes from various sources. For example, the federal government has placed a strong emphasis on student learning outcomes in college. The aforementioned Commission convened by the Department of Education points out a remarkable lack of data that allow meaningful comparisons across institutions (US Department of Education, National Center for Education Statistics 2006). Higher education institutions are called upon to adopt quality assessment instruments to measure student learning in order to improve instructional efficiency. In addition, accrediting organizations have increased the pressure on public colleges and universities to become more accountable for student learning as part of the accreditation process. Institutions are required to provide some form of a statistical report on student performance. More than 40 states have established an accountability or a statistical reporting system (Burke and Minassians 2002; Klein et al. 2005), and 27 states have a formal "report card" that characterize some of the 227 performance indicators that are related to student learning (Naughton et al. 2003). Finally, the rapid increase in college tuition and fees has provoked public inquiry about student learning in college. The published average annual tuition and fees for attending a four-year public university has almost tripled over the last three decades, from $2,303 in 1978 to $6,585 in 2008 (College Board 2008). Given the increasing financial cost of college education, public colleges and universities face mounting pressure to demonstrate to parents and other stakeholders that they have well utilized the investment and have well prepared students for careers beyond college.

Traditionally, institutions have been evaluated against many types of criteria, including graduation rate, student/faculty ratio, average admission test (i.e., SAT,[1] ACT) scores, the racial and ethnic composition of the student body, faculty resources and other human resources (Gates et al. 2001; Klein et al. 2005). In addition to demographics, student surveys also are commonly used in the evaluation of institutional quality. The surveys ask students about their learning experiences, level of engagement, satisfaction with their institutions, and career plans after college (Johnson et al. 1993). The actuarial data and student surveys provide useful information about the effectiveness and quality of an

[1] SAT and ACT are college admissions tests used in the United States.

institution. However, they do not provide any direct information on student learning. Quantitative results of student learning that allow direct comparisons across institutions have been requested by business leaders, faculty, accreditors, and community leaders (Nettles et al. 1997; Pascarella et al. 1996).

As a response to the public outcry for evidence of student learning that is comparable across institutions, many higher education organizations developed initiatives to provide institutions with the opportunities to demonstrate learning outcomes. For example, the Transparency by Design initiative, offered by WCET, a division of the Western Interstate Commission for Higher Education, aims to provide quality assurance on the learning outcomes reported by its member institutions. Transparency by Design is a groundbreaking initiative for institutions serving adult learners through distance education. Another example is the Voluntary System of Accountability (VSA) initiative. VSA was initiated by two leading organizations in higher education, the American Association of State Colleges and Universities (AASCU) and the Association of Public and Land-grant Universities (APLU; formerly known as the National Association of State Universities and Land-Grant Colleges [NASULGC]). VSA aims to measure core educational outcomes in college using standardized tests. Written communication and critical thinking were selected as the core educational outcomes as they represent the essential skills required for the twenty-first century global market.

VSA selected three standardized tests as its outcomes measures after reviewing a range of candidate measures. These three tests are the *ETS Proficiency Profile* (formerly known as the Measure of Academic Proficiency and Progress [MAPP], Educational Testing Service [ETS], 2007), the *Collegiate Assessment of Academic Proficiency* (CAAP; ACT 2009), and the *Collegiate Learning Assessment* (CLA; Council for Aid to Education 2007). These three measures were selected for several reasons: (a) these tests all include measures of written communication and critical thinking, (b) these tests have already been widely adopted as outcomes measures, thus making it easier for institutions to use these tests; and (c) these tests all have available research evidence showing adequate psychometric properties (ACT 2009; Marr 1995; Klein et al. 2007; Liu 2008, 2009b). VSA gives institutions the flexibility to choose any one of the three tests as their accountability measure.

Value-added computation

To measure instructional effectiveness in higher education, a term ‘value-added’ was introduced in VSA. Value-added is defined as the performance difference between first-year and fourth-year students on a standardized test (e.g., ETS Proficiency Profile, CAAP, CLA) after controlling for student admission scores (e.g., SAT, ACT) (Voluntary System of Accountability 2008). The value-added measure indicates how much students have learned in college in writing and critical thinking after taking into consideration their prior academic achievement. Institutions are then ranked based on their value-added scores.

Currently, VSA recommends a cross-sectional design for the value-added assessment. In a cross-sectional design, the groups of first-year and fourth-year students tested are not the same group of students. This design allows institutions to test the first-year and fourth-year students at the same time. Compared to the longitudinal design which requires tracking the first-year students for four years, the cross-sectional design is less costly and more feasible to implement. However, as VSA notes, if it becomes obvious that a longitudinal design is superior to a cross-sectional design, the decision on design selection will be revisited (Voluntary System of Accountability 2008).

To compute the value-added index, VSA recommends a regression method which is currently used with the CLA test (Council for Aid to Education 2007). The institution is used as the unit of analysis in the regression model. We use the ETS Proficiency Profile test as an example to illustrate this three-step method: (a) mean SAT score is used to predict mean ETS Proficiency Profile score in an ordinary least squares (OLS) regression model, and the mean is calculated at the institution level. This analysis is conducted for first-year and fourth-year students, respectively; (b) a residual score resulting from (a) is calculated for first-year and fourth-year students respectively at each institution; and (c) the final value-added index is determined by the difference between the residuals of first-year and fourth-year students. Each institution has such a value-added index and is rank ordered on the basis of the index. Specifically, the institutions are categorized into ten decile groups on the basis of the value-added index. For example, if an institution is placed in group 9, it suggests that this institution performed better, in terms of value-added, than 80% of the institutions included in the analysis. The three-step procedure is illustrated in Fig. 1.

As promising as the current method is in providing a benchmark to compare student learning gain across institutions, three methodological issues stand out. *First*, all of the student-level information is ignored using this method since the calculation is at the institutional level (Liu 2009a, b). Klein et al. (2008) also raised concerns about this method and proposed using the student as the unit of analysis for further research. The results could be more reliable using student-level information given that the number of students far exceeds the number of institutions in the equation. *Second*, the current method groups all types of institutions together for value-added analysis without considering any institutional characteristics. However, many factors, such as institution selectivity and type could have a profound impact on student learning, and should be considered in determining student progress (Liu 2008, 2009a, b; Borden and Young 2008). *Finally*, the current method uses OLS regression models to analyze student performance on outcomes tests. Given the hierarchical structure of the data with students nested within an institution, hierarchical linear modeling (HLM; Raudenbush and Bryk 2002) may be more appropriate than OLS models. One of the assumptions of the OLS models is that all of the observations are independent (Stone 1995). In the case of the value-added calculation for VSA purposes, it is very likely that student experience and learning are affected by the unique characteristics

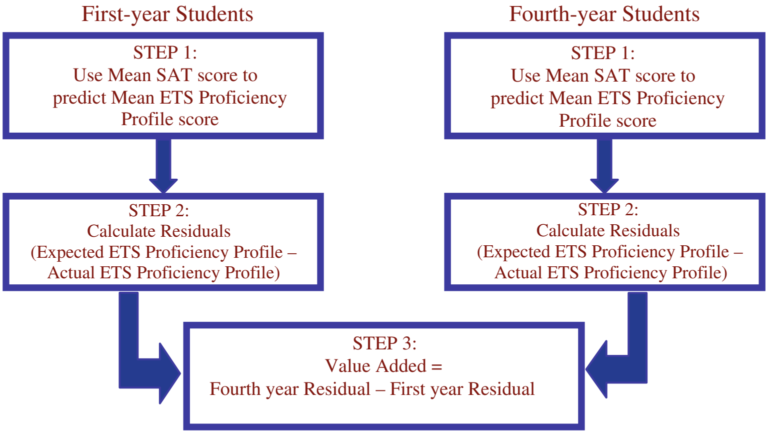
First-year Students
Fourth-year Students
STEP 1:
Use Mean SAT score to predict Mean ETS Proficiency Profile score
STEP 1:
Use Mean SAT score to predict Mean ETS Proficiency Profile score
STEP 2:
Calculate Residuals
(Expected ETS Proficiency Profile – Actual ETS Proficiency Profile)
STEP 2:
Calculate Residuals
(Expected ETS Proficiency Profile – Actual ETS Proficiency Profile)
STEP 3:
Value Added =
Fourth year Residual – First year Residual

Fig. 1 Three steps of current value-added computation in VSA

of the institution they attend. Therefore, test scores of students attending the same institution cannot be considered independent among themselves. Ordinary least squares analysis at the institution level ignores the interdependence among students who attend the same university and often results in an underestimate of the statistical significance of the institution-level effects (Lee 2000; Raudenbush and Bryk 2002). HLM is able to relax this constraint by differentiating the variance in student performance due to within-institution factors and between-institution factors (DiPrete and Forristal 1994; Raudenbush and Bryk 2002; Singer 1998).

Because of the relative recency of accountability research in higher education, not many studies exist investigating the use of value-added models. However, value-added research has been widely applied in K-12 settings. Most of the value-added investigations in K-12 are focused on the value introduced either by teacher (Lockwood et al. 2007; Yeh and Ritter 2009) or by school (Gorard 2008; Lee 2000; Palardy 2008). A shared purpose of the K-12 value-added research is to hold teachers, administrators, and schools accountable for student performance in school on a fair basis by considering differences in students' starting achievement. The Tennessee Value-Added Assessment System (Sanders and Rivers 1996; Sanders et al. 1997) was one of the most influential programs designed to evaluate the impact of teacher effectiveness. Based on results from about three million records of grade 2–8 students, Sanders and Rivers (1996) concluded that the differences in student achievement could be as large as 50 percentile points depending on the teacher a student had. Hierarchical modeling has been a commonly used tool in evaluating teacher or school effect given the nested structure of the data (e.g., students nested within a teacher, teachers nested within a school, schools nested within a district). Besides considering student starting performance, many research studies also control for students' socio-economic status (SES) and demographic backgrounds as these variables not only influence the starting point but may also affect student progress trajectories (Ballou et al. 2004; Linn 2001).

Objective of this study

This study aims to provide an alternative method for value-added computation. We then compare the institutional ranking resulting from the new method with the ranking from the OLS method used in VSA. To achieve this goal, this study reanalyzed data used in a research study conducted by Liu (2009b). The Liu (2009b) study examined how the ETS Proficiency Profile test should be used to evaluate institutional effectiveness and followed the OLS regression method recommended by VSA. In contrast, the new method considers important institution level predictors and takes advantage of multi-level modeling. When processing data with nested structure, multi-level models have well documented advantages compared to regression models (e.g., Luke 2004; Raudenbush and Bryk 2002; Richter 2006; Tekwe et al. 2004). This study uses a two-level hierarchical linear model in calculating value-added scores, with the first level being student and the second level being institution. Compared to the regression method currently in use with VSA, the HLM provides several benefits. First, it utilizes full information at the student level and presents a more accurate relationship between admission scores and standardized test performance. Since the current OLS regression analysis is conducted at the institution level and only considers mean scores, it may lead to an overestimate of the correlation between admission scores and standardized test scores. Second, the HLM method allows errors to exist at all levels of the data (e.g., student, institution) and can handle fixed or random effects, while

OLS regression assumes that errors only exist at the lowest level of the data. Finally, HLM accommodates the fact that students within an institution are not independent observations while the OLS regression treats all observations as independent. HLM's differentiation of both within- and between-institution variance leads to a more accurate estimate of variance in student performance. Although hierarchical models have not been widely used in value-added calculation in higher education, there are numerous examples of HLM applications in research about school or teacher effectiveness in K-12 (Lee 2000; Opdenakker and Van Damme 2006; Palardy 2008; Sanders and Rivers 1996; Yu and White 2002).

Methods

In this study, a value-added index was recalculated using the HLM method. Results from the HLM and OLS were compared in terms of institutional ranking. The following section details the ETS Proficiency Profile test, the sample used in this study, and the analysis methods.

The ETS Proficiency Profile test

The ETS Proficiency Profile was designed to measure college-level skills in critical thinking, reading, writing, and mathematics. It focuses on general academic skills developed through college education rather than focusing on the knowledge learned in specific courses. The ETS Proficiency Profile is offered in two forms: the standard form and a short form. The standard form consists of 108 items with 27 items in each of the four skill areas measured and takes 2 h to complete. The short form has 36 items and takes about 40 min to complete. All items are in multiple-choice format, and each item is situated in one of three academic contexts: humanities, social sciences, or natural sciences. ETS Proficiency Profile showed good internal consistency as indicated by Cronbach's alpha: .78 for critical thinking, .80 for writing, .81 for reading, and .84 for mathematics (Educational Testing Service 2007). Note that the reliabilities are calculated at the group level for institutions using the short form since for the short form students are required to respond to only about a third of the items included in the standard form. Both forms of the ETS Proficiency Profile are delivered via a paper/pencil format or via an online version. Scores from the two delivery formats are equated to make them comparable (Educational Testing Service 2007).

Students who take the standard form receive eight scaled scores, including a total ETS Proficiency Profile score, four skills subscores (critical thinking, reading, writing, and mathematics), and three content-based subscores (humanities, social sciences, and natural sciences). Similar information is provided at the group level for students who take the short form but not at the individual student level in order to achieve acceptable scale reliability.

The ETS Proficiency Profile also showed satisfactory predictive and content validity evidence. Based on data from more than 5,000 students, Marr (1995) reported that ETS Proficiency Profile scores increased as students advanced through the educational curriculum, with students who had completed all of the core curriculum credits scoring higher than students who had not. The finding provides evidence that student performance on the ETS Proficiency Profile test is not just a reflection of natural maturation. In addition, student performance on ETS Proficiency Profile skills was consistent with the skill requirements of their major fields of study, with humanities majors scoring higher than other students on critical thinking and writing, and mathematics and engineering students scoring higher on mathematics (Marr 1995).

Participants

To compare results yielded from both OLS regression and HLM, the same students examined in the Liu (2009b) study were included for analysis in this study. In total, 6,196 students participated in the Liu (2009b) study. These students came from 23 higher education institutions. They took the ETS Proficiency Profile test between 2006 and 2008. There was a large variation in terms of the incentives the institutions used to recruit students to take the test, including course credits, copy cards, cash, and book store coupons. Therefore, the sample of students attracted to take the test may differ from institution to institution. As a result, the degree to which the sample is representative of an institution may also vary from institution to institution.

After students took the ETS Proficiency Profile test, their SAT or ACT score was obtained from the institution registrar's office. Student ACT scores were transformed to be on the same scale as SAT scores using a concordance table (http://www.act.org/aap/concordance/). The concordance table is provided by the College Board and ACT as a tool for finding comparable scores between the two tests. It is reasonable to use admissions scores as a control for student starting achievement because the admissions tests measure similar constructs as the outcomes assessment. The SAT assesses critical thinking and problem solving skills in three areas: critical reading, mathematics and writing. Most of the test items use a multiple-choice format with some exceptions: the writing section uses a constructed-response format and some of the mathematics items use student-generated responses. The entire test takes three hours and 45 min (http://sat.collegeboard.com). The ACT measures knowledge and skills in four subject areas: English (75 questions), mathematics (60 questions), reading (40 questions), and science (40 questions). All of the questions are in a multiple-choice format. The entire test takes about three hours. ACT also has an optional writing section, which is an essay test and takes about 30 min to complete (http://www.actstudent.org/index.html).

When predicting the ETS Proficiency Profile scores using the admissions test scores, ideally the scale scores on the admissions tests should be used to predict the corresponding ETS Proficiency Profile scale scores (i.e., using SAT writing scores to predict ETS Proficiency Profile writing scores). However, because the concordance table between SAT and ACT is only available at the composite score level, the composite admission scores were used to predict ETS Proficiency Profile scores.

All of the students included in the study were full-time and non-transfer students as required by VSA. The participating students were comprised of 4,373 first-year students and 1,823 fourth-year students. The mean age was 18.42 (SD = .87) for the first-year students and 21.49 (SD = 1.21) for the fourth-year students. Note that the first-year students and fourth-year students were not the same group of students since it was a cross-sectional design. This sample consisted of 25.5% males, 45.5% females and the rest unknown. Six percent of the participants were Asian American, 16% African American, 60.5% White, 11% Hispanic, and the rest unknown (Liu 2009b).

Analysis

The Liu (2009b) study used the OLS regression method illustrated in Fig. 1. On the basis of the difference in score residuals between first-year and fourth-year students, each of the 23 institutions was ranked within ten decile groups. The ranking was assigned for both critical thinking and writing since these are the two skills emphasized by VSA. For purposes of comparison, the current study also focuses on critical thinking and writing. The method used in the current study for institutional ranking is described in this section.

Unconditional model. A two-level unconditional model was tested to determine whether a multilevel model is appropriate for the evaluation of institution value-added. Results from the model allow us to partition variance in student ETS Proficiency Profile scores into two components: within-institution variance and between-school variance. The ratio of between-school variance to the total variance (i.e., the sum of within- and between-institution variance) is referred to as the intraclass correlation coefficient (ICC). The ICC indicates how much variance is accounted by between-school factors.

The unconditional model did not include any predictors in either the Level 1 or Level 2 equations. It was tested for first-year and fourth-year students separately on ETS Proficiency Profile critical thinking and writing skills. The model is expressed as follow:

$$\text{Level 1: } Y_{\text{ij}} = \beta_{0\text{j}} + r_{\text{ij}} \tag{1}$$

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

where Y_{ij} is ETS Proficiency Profile critical thinking or writing score for student i at institution j. $\beta_{0\text{j}}$ is the mean score for all students in school j. r_{ij} is the score residual for student i at institution j, indicating the difference in ETS Proficiency Profile score between the student and school mean. γ_{00} is the grand mean score in ETS Proficiency Profile critical thinking or writing. $u_{0\text{j}}$ is institution level residual, indicating institution j's departure from the grand mean in terms of ETS Proficiency Profile scores.

Results from the unconditional model provided baseline information of the impact of institutions on student ETS Proficiency Profile performance. If the results show that institutions explain a considerable amount of variance (e.g., >10%), there is some evidence for using a multilevel model (Raudenbush and Bryk 2002; Lee 2000). The software HLM 6 was used to conduct the analysis (Raudenbush et al. 2004).

Intercept-as-outcomes model. In this model, at the first level, student admission score was used to predict their ETS Proficiency Profile score. Student ACT scores were transformed to the same scale as SAT scores using a concordance table (http://www.act.org/aap/concordance/). At the second level, school level characteristics including selectivity and degree-granting (i.e., with graduate programs or undergraduate only) were included as predictors to predict the first level intercept. Again, the HLM was analyzed for first-year and fourth-year students, respectively. The model is formulated as:

$$\text{Level 1: } Y_{\text{ij}} = \beta_{0\text{j}} + \beta_1 SAT_{\text{i}} + r_{\text{ij}} \tag{2}$$

$$\text{Level 2: } \beta_{0\text{j}} = \gamma_{00} + \gamma_{01}(\text{selectivity})_{\text{j}} + \gamma_{02}(DGI)_{\text{j}} + u_{0\text{j}}$$

where Y_{ij} is ETS Proficiency Profile critical thinking or writing score for student i in school j. $\beta_{0\text{j}}$ is the mean score for all students in school j. β_1 is the slope for the predictor of student SAT score. r_{ij} is the score residual for student i in school j. γ_{00} is the grand mean score in ETS Proficiency Profile critical thinking or writing. γ_{01} is the change in school mean score for one unit change in schools' selectivity. Selectivity is indicated by the percentage of students admitted among all applicants. DGI stands for degree-granting institution and was coded as a dummy variable with 0 indicating undergraduate programs only and 1 indicating graduate programs. The information on school selectivity and degree-granting was obtained from the 2008 College Handbook (College Board 2008). γ_{02} is the difference in school mean score between schools only offering undergraduate programs and schools with graduate programs. $u_{0\text{j}}$ is a random variable, indicating the difference between mean score for school j and the grand mean after controlling for the selectivity and degree-granting variables. $u_{0\text{j}}$ assumes a normal distribution with mean 0 and variance σ^2.

The intercept-as-outcomes model produces a residual estimate (u_{0j}) separately for the first-year and fourth-year students at each of the institutions. The difference in the residuals of the first-year and fourth-year students at a particular institution was used to indicate the value-added for that institution and the institutions were then ranked within ten decile groups on the basis of this value-added index. The procedure was repeated for both critical thinking and writing. Institutional ranking resulting from the multilevel models was compared to the ranking yielded from the OLS models.

Results

Descriptive statistics of the ETS Proficiency Profile critical thinking and writing are provided in Table 1. In general, the fourth-year students performed significantly better than the first-year students on both skills. Besides statistical significance, an effect size calculated by dividing the mean difference by the pooled standard deviation was used to indicate the magnitude of the performance differences. The effect sizes on both critical thinking and writing are considered large (i.e., >.80) according to the rules specified by Cohen (1988).

Table 2 presents the *Pearson* correlation between SAT scores and critical thinking and writing scores at the student level. We can see that SAT scores are moderately correlated with performance in these two ETS Proficiency Profile areas (Liu 2009b).

The random effects from the unconditional models are provided in Table 3. The results are organized by skill area—critical thinking and writing, and by class—first-year and fourth-year students. For the HLM models used in this study, it is important to test the homogeneity of variance assumption. If the variance of the level one residual r_{ij} is the same across all institutions, it indicates that institutions do not account for additional variance in student performance. Therefore, there would be no benefit of using a multi-level model. HLM 6 provides a chi-square test for homogeneity of variance. The significant chi-square values (Table 3) suggest heterogeneity of variance for first-year and fourth-year students on critical thinking and writing.

Table 1 Descriptive Statistics of the ETS Proficiency Profile Total Score, ETS Proficiency Profile Subscales, and SAT

	First-year students (n = 4,373)		Fourth-year students (n = 1,823)		t	d
	Mean	SD	Mean	SD		
Critical thinking	110	1.9	113	2.3	53.1***	1.4
Writing	113	1.7	115	1.8	41.7***	1.1

*** $p < .001$; d is the effect size of the differences between the first-year and fourth-year students

Table 2 Pearson correlations between ETS Proficiency Profile and SAT scores

	Student level	
	First-year	Fourth-year
Writing	.50	.59
Critical thinking	.54	.54

Table 3 Random effects from the unconditional models

	Random effect		SD	Variance component	*df*	Chi-square	*p*
Critical thinking							
First-year	Intercept	U_0	1.64	3.69	22	319.27	<.001
	Level 1	R	4.39	19.24			
Fourth-year	Intercept	U_0	2.10	6.39	22	195.58	<.001
	Level 1	R	6.12	37.49			
Writing							
First-year	Intercept	U_0	1.82	4.82	22	252.99	<.001
	Level 1	R	5.26	27.70			
Fourth-year	Intercept	U_0	1.66	3.96	22	221.23	<.001
	Level 1	R	4.37	19.08			

The intraclass correlation coefficient was calculated to indicate the percentage of variance explained by institutions among total variance. For example, the ICC for the first-year students on critical thinking in the unconditional model was 3.69/(3.69 + 19.24) = .16, which suggests that institutions accounted for about 16% of the variance in student ETS Proficiency Profile performance. Similarly, ICC was .14 for the fourth-year students on critical thinking, .15 for the first-year students on writing and .17 for the fourth-year students on writing. The percentages of school effects from the unconditional models are similar to findings from studies that investigate school effects in K-12 settings (Lee 2000; Palardy 2008). Also, all of the unconditional models showed good reliability (i.e., >.85) for the random intercept. The reliability of the first level intercept measures the ratio of the parameter variance to the total variance and is the average of the reliabilities across all units of the second level model. The findings support the use of a hierarchical model to disentangle the within- and between-institution effect.

Results of the fixed effects from the intercept-as-outcomes model are summarized in Table 4. At the student level, SAT was a significant predictor for ETS Proficiency Profile score for both skill areas and for both classes. At the school level, an institution's selectivity was a significant predictor of both critical thinking and writing scores for the fourth-year students, but not for the first-year students. Whether an institution offers graduate programs or not does not have a significant impact on student ETS Proficiency Profile performance for both first-year and fourth-year students. Results of the random effects of the intercept-as-outcomes HLM are summarized in Table 5. Again, the significant chi-square values point to the heterogeneity of variance across institutions in student outcomes performance, even after controlling for student SAT scores. The intraclass correlation was .09 for the first-year students on critical thinking, .10 for the fourth-year students on critical thinking, .09 for the first-year students on writing, and .11 for the fourth-year students on writing. The results consistently show that after student level and institution level predictors are included in the equations, the percentage of variance explained by institutions is reduced. The main reason was that given the high correlation between SAT scores and ETS Proficiency Profile scores, SAT scores accounted for a substantial amount of variance in ETS Proficiency Profile performance. The reliability of the student level random intercept β_{0j} was .72 for the first-year students and .74 for the fourth-year students on critical thinking, and was .81 for the first-year students and .78 for the fourth-year students on writing.

Table 4 Fixed effects from the Intercept-as-outcomes models

	Symbol	Coefficient	Standard error	*df*
Critical thinking				
First-year				
Student level				
SAT	β_1	.018**	.0005	4,369
Institution level				
Intercept	γ_{00}	91.7**	.772	20
Selectivity	γ_{01}	1.22	2.001	20
Degree granting	γ_{02}	.522	.474	20
Fourth-year				
Student level				
SAT	β_1	.022**	.0008	1,819
Institution level				
Intercept	γ_{00}	89.05**	1.056	20
Selectivity	γ_{01}	.53*	.27	20
Degree granting	γ_{02}	.16	.32	20
Writing				
First-year				
Student level				
SAT	β_1	.014**	.0004	4,369
Institution level				
Intercept	γ_{00}	99.21**	.871	20
Selectivity	γ_{01}	.89	.65	20
Degree granting	γ_{02}	.12	.44	20
Fourth-year				
Student level				
SAT	β_1	.015**	.0009	1,819
Institution level				
Intercept	γ_{00}	100.1**	1.13	20
Selectivity	γ_{01}	.52*	.26	20
Degree granting	γ_{02}	.57	.38	20

* $p < .05$; ** $p <.01$

The outcome that institutions are probably most interested is their value-added ranking. As described in the method section, the difference between the freshman and senior residual (u_{0j}) at each institution was used to indicate final value-added in this study. And the rankings were compared with the rankings produced from the ordinary least squares method recommended by VSA. Figures 2 and 3 illustrates the two sets of institutional rankings on critical thinking and writing, respectively. The correlation between these two sets of rankings was .76 for critical thinking and .84 for writing. Although the rankings produced by the different procedures are highly correlated, the impact of the difference in the ranking could be significant for a particular institution. Among the 23 institutions included in the analysis, only 5 institutions were ranked the same on critical thinking using both methods, and only 8 institutions were ranked the same on written communication. The

Table 5 Random effects from the intercept-as-outcomes models

Random effect		SD	Variance component	*df*	Chi-square	*p*
Critical thinking						
First-year						
Intercept	U_0	.79	2.05	20	214.13	<.001
Level 1	R	4.48	20.09			
Fourth-year						
Intercept	U_0	.98	3.04	20	62.34	<.001
Level 1	R	5.12	26.23			
Writing						
First-year						
Intercept	U_0	.55	1.56	20	162.58	<.001
Level 1	R	3.86	14.91			
Fourth-year						
Intercept	U_0	.63	1.88	20	81.71	<.001
Level 1	R	3.82	14.58			

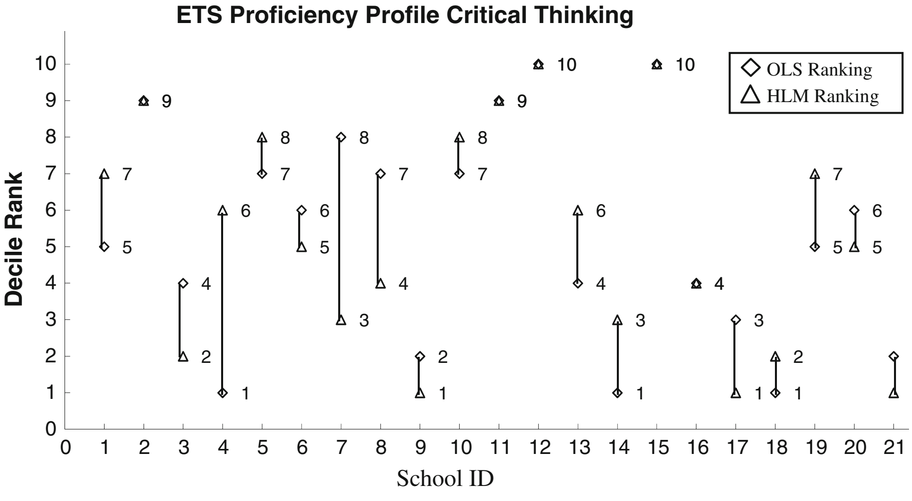

Fig. 2 Value-added ranking in critical thinking using the OLS method and the HLM Method

differences in ranking can be as large as 5 levels out of the ten decile groups on critical thinking (e.g., from decile ranking 1 to 6 or from 3 to 8). Similarly, the change in ranking can also be substantial on writing. The largest difference in ranking on writing went from decile group 1 to group 5 for one institution, and from 3 to 7 for another.

Discussion

As college tuition and fees continue to increase, stakeholders at many levels are interested in understanding how public investment is utilized and what can be done to improve

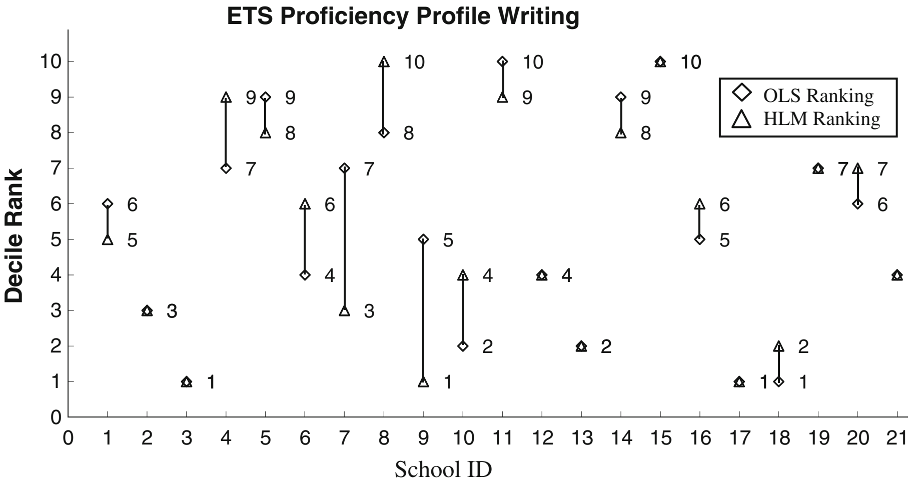

Fig. 3 Value-added ranking in writing using the OLS method and the HLM method

institutional efficiency. Institutions have been put under the spotlight and the evaluation of their effectiveness has received unprecedented attention. As a response to public inquiry, the Voluntary System of Accountability was initiated and has offered a method to calculate student learning gain from first year to fourth year. After controlling for admission score, the performance difference between first-year and fourth-year students on a standardized test is referred to as the value-added. As promising as the current method is in comparing student progress across institutions, it has some potential limitations lying in its unit of analysis and the statistical model it employs. Given that VSA is a relatively recent development, additional research is needed to evaluate alternative methods in order to identify a meaningful and accurate way of calculating value-added for each institution.

The method used in this study in determining institutional value-added differs from the currently method used for purposes of VSA in three aspects. First, the student was used as the unit of analysis at level one of the multilevel models. This way, we are able to accurately estimate the relationship between student admission scores and their performance on the ETS Proficiency Profile test. Second, two institution level predictors, selectivity and degree-granting of an institution, were considered in the evaluation of school impact on student ETS Proficiency Profile performance. Between these two variables, selectivity was found to be a significant predictor of ETS Proficiency Profile scores in both critical thinking and writing for senior students. The finding provides evidence that important institution level factors should be considered in the value-added calculation. Finally, this study used hierarchical linear models instead of OLS regression models in deriving value-added scores for each institution. As discussed in previous sections, HLM has many well-documented advantages over OLS models (DiPrete and Forristal 1994; Raudenbush and Bryk 2002; Singer 1998).

Results show that the institution rankings are highly correlated using the HLM and OLS models, for both critical thinking ($r = .76$) and writing ($r = .84$). However, the difference in ranking may have significantly different consequences for certain institutions. For example, a university was ranked at the lowest level (i.e., decile group one) on critical thinking skills using the OLS method. Its ranking increased to decile group 6 after the HLM was applied to the same data, which means that using the new calculation this university had achieved more value-added in critical thinking than 50% of the institutions included in the comparison. The conclusion about the effectiveness of this institution will be diametrically different depending upon which method is used to calculate value-added scores. In addition, should the value-added results ever be considered for inclusion in accreditation process, decisions about the quality of that institution will also be very different depending on which value-added method is used. Therefore, it is critically important to identify a meaningful and accurate method for calculating value-added score in order to provide a fair evaluation of institutions.

Besides methodological issues, there are also other notable challenges in determining institutional effectiveness in current VSA. As institutions have the flexibility to choose from the three tests (i.e., ETS Proficiency Profile, CAAP, and CLA) as their outcomes test, evidence is urgently required on the comparability of results from the three tests. An institution's value-added ranking could vary significantly depending on which test it selects to use. To address this issue, the testing organizations that sponsor the three tests, ETS, ACT, and CAE are working on a construct validity study to investigate the relationship among measures designed to assess similar constructs on the three tests. For example, they are interested in finding out the correlation between ETS Proficiency Profile critical thinking and CLA critical thinking. They are also interested in investigating whether item format has an impact on student performance on these standardized tests, since both ETS Proficiency Profile and CAAP are multiple choice item tests and CLA is an essay type test. This study is funded by the Department of Education under the Fund for the Improvement for Postsecondary Education program.

For future research, more attention should be devoted to advancing the value-added model when determining institutional effectiveness. Student learning and growth in college is influenced by many individual and institutional factors, and these factors should be considered for an effective evaluation of institutional efficiency. The resources that institutions have access to should also be factored in when determining whether colleges and universities have done an adequate job preparing students with the skills needed for the twentyfirst century. Although research on value-added assessment is relatively new in higher education, abundant research has been conducted on value-added for K-12 schools. For example, in K12 settings, student characteristics such as gender, language status, ethnicity, and socio-economic status have been shown to be strong predictors of student achievement and learning (Palardy 2008; Park and Palardy 2004). In the context of higher education, as students have more flexibility and freedom, their motivation to learn can also have a significant impact on their learning outcomes (Rodgers 2007). Students' major fields of study may also influence their performance on general outcomes assessments. For example, science and engineering students may have advantages on the mathematics test (Marr 1995).

At the institutional level, institutional resources in terms of both fiscal resources and human resources, such as student to faculty ratio, degree-granting status, and selectivity could also have an impact on student progress (Darling-Hammond et al. 2001; Rodgers 2007). Therefore, these factors could be considered when evaluating value-added of colleges and universities. What's more, as results from research on K12 schools show that student learning and experiences are very different in different types of schools (i.e., public, Catholic, private schools) (Lee 2000; Palardy 2008), probably the same distinction should be drawn when comparing higher education institutions.

The current study only considers student admissions scores at the student level model. Data on student learning experience and other demographic variables are not available and therefore are not included in the analysis. At the school level, only institutional selectivity and degree-granting status are considered for parsimony of the model as only 23 institutions are available at the second level regression equation. These are the limitations of this study. Future research should expand the scope of the investigation by including more universities to the sample and factor in more student- and institution-level variables to obtain a more complete picture of how these variables affect student learning outcomes. Future research should also compare institutions with similar student bodies instead of grouping institutions of all kinds for value-added ranking. For instance, it may not be fair to compare student learning at a local liberal arts college to learning at a resourceful national university.

Finally, a logical next step would be to experiment with the longitudinal design to determine to which degree the cross-sectional results can approximate the longitudinal results. Although the current cross-sectional design has great implementation feasibility, no research evidence exists supporting the similarities in results between the cross-sectional and the longitudinal design. If results from the two kinds of designs show considerable differences, we need to reconsider the choice of study design.

Results from this study have demonstrated the need for additional research in the evaluation of institutional effectiveness for accountability purposes. It is important to consider the context of learning while comparing student performance on standardized tests across institutions. The notion of value-added has been a controversial topic among higher education researchers and institutions. Although institutions have significant responsibilities for student learning, it is also influenced by many other factors such as student motivation, academic engagement, college readiness, and career aspirations. These factors are largely out of an institution's control. As the results of value-added ranking will likely have a profound impact on institutions, we caution stakeholders to be careful in interpreting current value-added scores and in establishing a link between student learning and institutional efficiency.

Acknowledgments I want to thank Paul Holland and Sandip Sinharay for discussions with them during the development of this research project. I also want to thank Yue (Helena) Jia for her helpful suggestions on the use of the hierarchical models.

References

ACT. (2009). *CAAP guide to successful general education outcomes assessment*. IOWA City, IA: ACT.

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.

Borden, V. M. H., & Young, J. W. (2008). Measurement validity and accountability for student learning. In V. M. H. Borden & G. R. Pike (Eds.), *Assessing and accounting for student learning: Finding a constructive path forward*. San Francisco: Jossey-Bass.

Burke, J. C., & Minassians, H. (2002). *Performance reporting: The preferred ''no cost'' accountability program (2001)*. Albany: The Nelson A. Rockefeller Institute of Government.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

College Board. (2008). *College Handbook 2008*. The College Board: New York.

Council for Aid to Education. (2007). *CLA institutional report 2006–2007*. New York: Council for Aid to Education.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*, 57–77.

DiPrete, T. A., & Forristal, J. D. (1994). Multilevel models: Method and substance. *Annual Review of Sociology, 20*, 331–357.

Educational Testing Service. (2007). *MAPP user's guide*. Princeton, NJ: ETS.

Gates, S. M., Augustine, C. H., Benjamin, R., Bikson, T. K., Derghazarian, E., Kaganoff, T., et al. (2001). *Ensuring the quality and productivity of education and professional development activities: A review of approaches and lessons for DoD*. Santa Monica, CA: National Defense Research Institute, RAND.

Gorard, S. (2008). The value-added of primary schools: What is it really measuring? *Educational Review, 60*(2), 179–185.

Johnson, R., McCormick, R. D., Prus, J. S., & Rogers, J. S. (1993). Assessment options for the college major. In T. W. Banta, et al. (Eds.), *Making a difference: Outcomes of a decade of assessment in higher education* (pp. 151–167). San Francisco: Jossey-Bass.

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*, 511–525.

Klein, S., Kuh, G., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher-education institutions. *Journal of Research on Higher Education, 46*(3), 251–276.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125–141.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems. CSE Technical Report 539*. Los Angeles: University of California.

Liu, O. L. (2008). *Measuring learning outcomes in higher education using the Measure of Academic Proficiency and Progress (MAPP^{TM}). ETS Research Report Series (RR-08–047).* Princeton: ETS.

Liu, O. L. (2009a). *Measuring learning outcomes in higher education (Report No. RDC-10).* Princeton, NJ: ETS.

Liu, O. L. (2009b). Measuring value-added in higher education: Conditions and caveats. Results from using the Measure of Academic Proficiency and Progress (MAPP$^{\mathrm{TM}}$). *Assessment and Evaluation in Higher Education, 34*(6), 1–14.

Lockwood, J. R., McCaffrey, D. F., & Hamilton, L. S. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Marr, D. (1995). *Validity of the academic profile*. Princeton, NJ: ETS.

Naughton, B. A., Suen, A. Y., & Shavelson, R. J. (2003). Accountability for what? Understanding the learning objectives in state higher education accountability programs. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Nettles, M. T., Cole, J., & Sharp, S. (1997). A comparative analysis of trends and patterns in state assessment policy. Paper presented at the annual meeting of the Association for the Study of Higher Education Conference, Albuquerque, NM.

Opdenakker, M. C., & Van Damme, J. (2006). *Principles and standards for school mathematics*. Reston, VA: Opdenakker MC & Van Damme J.

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multilevel, multiple group latent growth curve analysis. *School Effectiveness and School Improvement, 19*, 21–49.

Park, E., & Palardy, G. J. (2004). The impact of parental involvement and authoritativeness on academic achievement: A cross ethnic comparison. In S. J. Paik & H. Walberg (Eds.), *Advancing educational productivity: Policy implications from national databases* (pp. 95–122). Greenwich, CT: Information Age.

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*, 159–174.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*(3), 221–250.

Rodgers, T. (2007). Measuring value added in higher education: A proposed methodology for developing a performance indicator based on teachers economic value added to graduates. *Education Economics, 15*(1), 55–74.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Tennessee: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Education and Behavioral Statistics, 24*(4), 323–355.

Stone, C. (1995). *A course in probability and statistics*. Belmont, CA: Duxbury Press.

Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36.

U.S. Department of Education, National Center for Education Statistics. (2006). *Digest of Education Statistics, 2005* (NCES 2006-030).

Voluntary System of Accountability. (2008). *Background on learning outcomes measures.* Retrieved May 18, 2009, from www.voluntarysystem.org/index.cfm?page=about_cp.

Yeh, S. S., & Ritter, J. (2009). The cost-effectiveness of replacing the bottom quartile of novice teachers through value-added teacher assessment. *Journal of Education Finance, 34*(4), 426–451.

Yu, L., & White, D. B. (2002). Measuring value-added school effects on Ohio six-grade proficiency test results using two-level hierarchical linear modeling. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans: LA.