Using Student Learning as a Measure of Quality in Higher Education*

Stephen R. Porter †

November 22, 2011

*Unpublished working paper developed for HCM Strategists, LLC. Do not cite or redistribute without permission.

†Address: Department of Leadership, Policy, and Adult and Higher Education, Box 7810, North Carolina State University, Raleigh, NC 27695. E-mail: srporter@ncsu.edu. I would like to thank Sandy Baum, Kristin Conklin, Sylvia Hurtado, Nate Johnson, John Pryor and Jeffrey Steedle for comments. This research was supported by the Gates Foundation through their Context for Success project.

1 Introduction

Accountability for colleges and universities has greatly increased during the past several decades, driven by stakeholder interest in understanding exactly what institutions of higher education are accomplishing. At the same time, consumer demand for information about college quality has fueled a mini-industry of commercial college rankings. Yet there is common agreement among most educational researchers and policymakers that quality measures based on student inputs, such as SAT scores, tell us little about institutional quality. And while degree completion rates provide one way to compare institutional quality, they tell us little about the quality of the *education* that students receive.

Given that students attend college to increase their human capital, it is difficult to discuss quality metrics for higher education without some consideration of student learning. In the end, receiving a diploma for completing a certain number of credit hours is not what matters; it is instead what students have learned while earning those credit hours. Logically, students who attend college should greatly increase their knowledge, skills and abilities during their studies. Yet recent research suggests that many students learn very little during their years at college.[1]

The purpose of this paper is to review existing measures of student learning, and to explore their strengths and weaknesses as a quality metric for higher education. Unlike the K-12 arena, which employs standardized testing in a variety of subject areas, there is little consensus on how we should measure learning in higher education.

Considering the current discussion about higher education in the U.S., the idea of quality is implicitly a comparative one: Students, families, stakeholders and policymakers wish to distinguish between high- and low-quality institutions. Thus, any measures of student learning that we might consider must be comparable across some institutions and, we hope, across a wide variety. This approach rules out the use of grade-point averages or portfolios as useful national measures of student learning. Grading approaches differ widely across institutions, as do the requirements for portfolios. Evaluating portfolios in a consistent manner across institutions also appears to be a formidable task, given the time and judgment required.

Instead, I focus on three approaches currently used by many colleges and universities: 1) student self-reports of measures, such as frequency of contact with faculty and hours spent studying, that are believed to be highly correlated with student learning; 2) student self-reports of their learning gains during their time in college, in areas such as critical thinking and quantitative skills (the National Survey of Student Engagement (NSSE) and surveys produced by other providers are examples of 1 and 2); and 3) exam-based measures of student learning, where students are tested as to the level of their knowledge or broader skills such as critical thinking (e.g., the Collegiate Learning Assessment (CLA)).

I argue that any measure of student learning used for institutional comparisons will be valid only if three conditions hold. First, the measure must be very strongly correlated with student learning; that is, it must be valid at the individual level. This may seem an obvious point, but as this review will show, some commonly used measures do not appear related to student learning. Second, students for whom we have measures at an institution must be representative of all students at the institution. If not, then inter-institutional comparisons become difficult, as the subset of participants may be very different from the typical student at an institution. Third, the comparisons must take into account the heterogeneity of student ability among institutions. Not surprisingly, students at more selective institutions score higher on assessments than those at less selective institutions. If metrics do not take into account the different "starting places" of institutions, then more selective and wealthy institutions may be rated as higher quality, not because they do a better job of teaching students, but simply because of their admissions practices and the makeup of their student body. It is difficult to justify measuring institutions on a learning-based metric if these three conditions cannot be met, because some institutions will invariably be penalized because of methodological issues, rather than learning issues.

This review of the literature should be useful for institutional policymakers and governance bodies (e.g., university system offices, boards of trustees and accreditors), as well as grant agencies and foundations seeking postsecondary research areas to fund. For institutional policymakers, the choice of measure is an important one. Measures such as the NSSE and CLA

promote the use of benchmarks, based on the performance of other institutions. If the measures are flawed, then internal comparisons to these benchmarks may result in the misallocation of resources to unneeded programs, while steering resources away from more effective programmatic efforts. For university system officials and accreditors, it is vital that any metric chosen for evaluating institutions be an accurate measure of student learning at *all* institutions under consideration. If not, some institutions may be unfairly penalized for what seems to be low performance, while others are unjustly rewarded for seemingly high performance. For research funders, it may not be immediately obvious where funds should be spent to advance our ability to accurately assess student learning, given the array of learning assessments and numerous studies promoting their virtues.

2 What do we mean by student learning in college?

Before we can evaluate measures of student learning, we need a definition of what we mean by *learning*. As noted by Arum and Roksa, there is widespread agreement that developing critical thinking is one of the primary purposes of college.[2] By critical thinking, most observers refer to the ability to engage in the "... process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action."[3] While students are expected to develop specific knowledge and understanding of their academic major discipline, the ability to engage in critical thinking should occur across all fields, thus providing an appropriate inter-institutional measure of student outcomes. In addition, the concept of learning typically focuses on growth during college: namely, that students should be better at critical thinking when graduating from college compared with when they entered.

While there are a wide variety of college student learning assessments, ranging from knowledge tests of specific fields to measures of students' verbal and quantitative skills, I focus on critical thinking in this review for two reasons. First, most observers would agree that criti-

cal thinking can be considered the meta-measure of student learning; that is, we might wish for students to accomplish much during college, but critical thinking would be at the top of the list. Second, the number of student learning assessments is quite large, and beyond the scope of a single paper to evaluate. I note, however, that many of the issues raised by this review also apply to these other assessments.

3 Do common assessments measure learning in college?

This section reviews existing research on common assessments of learning to answer a simple question: To what extent do these assessments measure (or correlate with) student learning?

3.1 Self-reports of behavior

Probably the most common measure of student learning is based on scales constructed from self-reports of behaviors thought to be highly correlated with student learning. These behaviors include such activities as frequency of contact with faculty, frequency of discussions with other students, and hours spent on a variety of activities, including studying and cocurricular activities. The National Survey of Student Engagement is the most widely known survey that uses this approach, often referred to as measuring a construct termed "student engagement." Other surveys that attempt to measure student behavior include a survey similar to the NSSE for community college students (CCSSE), surveys produced by the Higher Education Research Institute (HERI) at UCLA, as well as many surveys designed and administered by individual institutions.[4]

These self-reported behaviors have several drawbacks that prevent their use as proxies for student learning. While critics have noted several shortcomings with these questions, such as vague wording, the most problematic is the ability of students to accurately report on their behavior.[5] Much of what we seek to know about student engagement is in the realm of mundane and frequent activities. Students are asked, for example, how often during the current academic year

they asked questions in class or contributed to class discussions, came to class without completing readings or assignments, and failed to complete homework and course assignments on time.[6] For many students, these activities could happen every week, or even on a daily basis.

Memory research demonstrates that individuals have difficulty in reporting on these kinds of activities; instead, it is infrequent and unusual experiences that we can easily report.[7] This memory problem is compounded by the reporting period most survey questions use. The NSSE asks students to report on an entire academic year, while the HERI survey of seniors asks about frequencies of behavior since entering college. Research indicates that individuals report behaviors most accurately when questioned near the time of the reported event, not weeks or months afterward.

Hours spent on activities thought to be beneficial or detrimental to learning, such as preparing for class, participating in cocurricular activities and socializing, are a second set of commonly asked survey questions thought to be correlated with learning.[8] The same memory issues listed above occur with these types of questions as well. A large body of research, dating back several decades, demonstrates that respondents are unable to accurately report on how they spend their time, unless asked about the previous 24-hour period.[9] Instead, researchers use time-use diaries that query respondents about their activities in the previous 24-hour period. Notably, this is how the federal government collects information about time use in the Bureau of Labor Statistics American Time Use Survey.

How exactly, then, do students provide answers to questions about frequency of behavior and time spent, if they cannot accurately recall and report this information? While survey response rates are declining, and it is becoming increasingly difficult to persuade students to respond to survey requests, many educational researchers and policymakers are not aware of a surprising fact: Those who agree to fill out surveys are more than happy to provide answers to questions, even if it is theoretically impossible to do so. For example, research indicates it is not unusual for respondents to provide answers to questions about fictitious issues.[10]

Some scholars have theorized that college students use context to generate responses.[11]

For example, a recent study provided random samples of students with two versions of the NSSE, one with a low scale (ranging from "0" to "4 or more times" in a given time period) and one with a high scale (ranging from "less than 4" to "10 or more times"). The proportion of students reporting engagement behavior four or more times differed on average by 25 percentage points between the two samples, simply because of the small change in the response scale.[12] Unable to answer the question based on their recall, many students used the response scale to generate a response, reasoning that middle responses corresponded to the typical student.[13] Students may use the reputation of their university (grind versus party) or aspects of their college major ("If I am a major in the natural sciences, I must be doing certain things") to generate a response, independent of their actual behavior. Research also suggests that social desirability bias plays a role in student responses. Students tend to over-report on items that make them look good, such as grade-point average, and under-report on negative information, such as being on financial aid. [14]

Given these issues, we would not expect to see strong correlations between self-reported behavior and student learning. Thus, it is not surprising that studies using exam-based measures of student learning and critical thinking find remarkably small relationships between learning and how high a student scored on a variety of scales derived from the NSSE.[15] Almost all of the effect sizes in these studies are less than .10, and the majority are equal to zero. Research also finds almost no relationship between NSSE scales and postsecondary outcomes such as college grade-point average and persistence.[16] Given that the use of self-reports of behavior as a proxy for learning requires a high correlation between the two, this body of research suggests that self-reports are useful neither as a measure of institutional learning nor as a measure of institutional quality in general.

In sum, self-reports of behaviors and time spent rely on an unrealistic view of how students respond to surveys: namely, that students have a computer hard drive in their brain that allows them to accurately recall and report on a wide variety of mundane and frequent activities that may be of minimal interest to them. Research seeking to link these data to objective measures of learning and critical thinking, and measures of student success such as GPA and persis-

tence, have found little to justify their use as proxies of student learning.

3.2 Self-reports of learning gains

A second set of self-reports have also been used as measures of learning during college. Here, students are asked a question regarding the amount of change they have experienced during college in a wide variety of content and skills areas, including critical thinking. Students are asked this question at various time points during their academic career, usually at the end of their first year of classes and at the end of their senior year.

The commonly accepted model of survey response posits that students must first comprehend the question, recall the relevant information from their memories, use this information to create an answer, and then map their answer onto the response scale on the survey.[17] For self-reports of learning gains, the cognitive effort required for accurate response is immense. Students must know their level of knowledge in a specific content area at college entry on some unspecified scale of knowledge, remember these levels one to six years later, know their level of knowledge in the area when administered the survey, be able to calculate the difference between the two and, finally, map their estimated differences onto the vague quantifier response scale.[18]

Given the memory issues discussed above, it is doubtful that most students have the cognitive ability to report on how much their learning has changed over the course of their college experience, and scholars using these questions have not offered any theoretical explanation as to how accurate responses might be possible. Not surprisingly, empirical research in this area supports the theoretical prediction that students are not able to provide accurate responses to self-reported gains questions.

Using exam-based measures of critical thinking and moral reasoning at the beginning and end of the school year, actual changes have been matched to self-reported changes in these two areas. The correlations between the two measures of change have been very low, often zero.[19] Given that the two measures are attempting to measure the same construct, we would expect the correlations to be close to 1. Theoretical models of cognition, and empirical evidence to date,

demonstrate that self-reported learning gains are mostly noise and cannot be used to assess student learning.

3.3 Exam-based measures of learning

Exam-based measures of learning are increasingly popular, as evidenced by the publicity surrounding the Collegiate Learning Assessment. Such assessments attempt to measure student learning directly, as opposed to measuring attitudes and behaviors thought to be correlated with learning. I refer to these as exam-based, as the administration of the assessment is quite different from surveys. Exam-based measures usually require a proctored administration, in which students are timed while taking the assessment. The most popular of these assessments are the Collegiate Assessment of Academic Proficiency (CAAP; produced by ACT), the Collegiate Learning Assessment (CLA; produced by the Council for Aid to Education) and the Proficiency Profile (PP, referred to as the Measure of Academic Proficiency and Progress (MAPP) before 2009 and Academic Profile before 2006; produced by ETS). These are also the three direct assessments recognized by the Voluntary System of Accountability.

The difference between these assessments is how they measure critical thinking. Both the CAAP and PP give students several different readings, followed by multiple-choice questions asking students to evaluate claims made in the readings and to evaluate statements about specific aspects of the readings. Based on sample test questions available to the public, the PP uses passages of 3–5 sentences in length, while the CAAP uses longer passages of 30–50 sentences. The PP examples are short reading passages, poems, graphs or tables, while the CAAP uses a variety of formats, such as presenting two sides of a debate, a dialogue between individuals arguing over an issue, case studies, statistical arguments, experimental results and editorials.[20]

The CLA Performance Task provides students with several related readings and asks them to write an essay evaluating some aspect of the readings. In the DynaTech company airplane example, students are given several artifacts: a newspaper article, a federal accident report on airplanes, company e-mails, charts on airplane performance, a magazine article and information

about two plane models. Then they are asked to write a memo evaluating the safety of a specific model of plane, conclude whether the plane should be purchased, and justify the recommendation. Each artifact is roughly the same length as the CAAP examples.

If we extend the definition of critical thinking listed above to include *multiple* sources of information, then the CLA would seem to have a strong claim to content validity. The CLA Performance Task requires students to analyze and evaluate a wide variety of related readings, then synthesize these readings in the response essay. However, given the nature of the response task, it is clear that the Performance Task is simultaneously measuring both critical thinking as well as writing skills: An excellent critical thinker could score low simply because of poor writing.[21] While the CLA has attracted wide attention because of its use of multiple information sources, the dual nature of the instrument is somewhat troubling, because the score is measuring two different constructs.[22]

The CAAP also has strong content validity, because it requires students to analyze and evaluate a wide variety of readings, albeit unrelated readings. Given the nature of the response (multiple-choice), students are not given the ability to synthesize. How important this is depends on one's definition of critical thinking: Is evaluating competing claims more important than synthesizing a group of related readings? Unlike the CLA, the CAAP (as well as the PP) appears to measure only one construct, critical thinking, because of the use of a multiple-choice response.

The PP has the least claim to content validity, simply because the passages students are asked to evaluate are short and do not appear to be particularly challenging. The short passages, followed by one or two questions, do not force students to critically evaluate a long, complex argument the way both the CAAP and CLA do. The CAAP, for example, provides students with a dialogue between two individuals approximately 50 sentences long, and then asks five questions about the material. The CLA provides students with an even larger set of material. Grappling with large, complicated sets of information is generally what many people think about when conceptualizing critical thinking, not answering single questions about a few sentences.

My review of the literature reveals the following evidence for criterion validity: that is,

the extent to which the three instruments are related to other external measures, such as GPA. For the CLA, studies indicate that the Performance Task performance is correlated with SAT scores (.54 to .56) and GPA (correlations range from .50 to .72); correlations using institutional-level means are higher (.65 to .92).[23] The MAPP (precursor to the PP) shows a similar pattern, with a correlation of .54 with SAT at the student level, .85 to .88 at the institutional level, and a positive relationship with GPA.[24] MAPP scores also vary by major field, with humanities, science, and engineering majors scoring higher than majors in business and education.[25] Arum and Roksa find a similar pattern for the CLA.[26] Research by ACT shows that the CAAP is correlated with GPA at the individual level (correlations range from .26 to .35).[27]

More work has been done in the area of construct validity, determining whether measures of critical thinking vary with similar measures. The Test Validity Study is probably the most comprehensive validity study to date of the three instruments.[28] Using data from over 1,100 students at 13 schools, all three instruments were administered, allowing comparisons at both the student and institutional levels. Student-level correlations between the measures are somewhat problematic, because the CLA is not designed to be a reliable measure at that level; students are not tested on multiple items as with the CAAP and PP. The school-level correlations are more comparable, and reveal something quite interesting: Schools with students that do well on one measure also do well on the other measures (see Table 1). At the student level both the CAAP and MAP are highly correlated. Together, these correlations suggest that the three measures are measuring the same construct.[29]

Table 1: Correlations between the CAAP, CLA and MAPP

	Student-level		School-level	
	CLA	MAPP	CLA	MAPP
CAAP	.47	.75	.79	.93
CLA	-	.53	-	.83

Another component of construct validity is growth.[30] For any measure of student learning to be considered valid, substantial first-year to senior differences should occur. Longitudinal

studies of the CLA Performance Task report effect sizes for growth that range from .18 to .40.[31] The .18 estimate is the increase after two years, while the .40 estimate is the increase after four years of college. Growth rates for the CAAP from the most recent Wabash longitudinal study are similar: .11 after the first year and .44 after four years.[32] Similar longitudinal measures are not available for the PP.

Other studies have compared first-year students and seniors at the same point in time, the idea being that seniors should score higher than first-years because of growth in college. It should be emphasized that these are not true growth estimates; instead, it is assumed that seniors' levels of critical thinking when they entered college are equal to the level of current first-year students. The Test Validity Study found effect sizes for the two groups of students to be .23 for the CLA Performance Task, .31 for the CAAP, and .46 for the MAPP (these are corrected for the difference in SAT scores between first-years and seniors).[33] Another CLA study found a much larger effect size of around 5.0 for the Performance Task.[34] Additional research on the MAPP demonstrates a very large increase of 1.4 standard deviations from first-years to seniors.[35] However, this effect size is somewhat exaggerated, as the first-year sample includes students who would drop out of school and not appear in the senior sample in a true longitudinal design. The difference in SAT scores was .57 SD, suggesting that much of this very large effect was due to attrition, and that the adjusted effect size would be more in line with estimates from other studies.

Whether these should be considered substantial is open to debate; the question is determining how much of an increase in critical thinking colleges can actually achieve. One possibility is to look at the K-12 literature on the effect of educational interventions on student achievement. A meta-analysis of average effect sizes estimated by other meta-analyses found mean effect sizes in the .20 to .30 range. Notably, effect sizes from randomized studies drop dramatically as the outcome measure moves from specialized topics (.44) to narrow standardized tests (.23) to broad standardized tests (.07).[36] These results suggest that changing student performance on broad measures such as critical thinking is difficult, and that the four-year changes in these instruments are comparable to many effects found in the K-12 literature. In other words, these instru-

ments are measuring *substantively* significant first-year to senior year growth.

These large changes over time are even more impressive when we consider student motivation. A large literature shows that student motivation has an impact on low-stakes test performance, such as the critical thinking instruments reviewed here (the term "low-stakes" refers to the stakes for the students; there are few, if any, penalties for poor performance that matter to the student, unlike state high school graduation exams and the SAT).[37] Using a value-added approach with measures of critical thinking at entry and exit, student motivation will not matter if the level of motivation for a student remains constant between the two testing periods. My personal experience with survey response rates is that seniors tend to be more jaded than first-year students, suggesting that their motivation during low-stakes exams is lower, which in turn implies that their critical thinking scores are lower than they would be if their motivation was similar to first-year students. If generally true, this suggests the growth estimates cited above are underestimates of how much students learn in college.

3.4 Summary

While both theory and research indicate that student self-reports are not useful correlates of student learning, a growing body of evidence demonstrates that the CAAP, CLA and PP are valid measures of student learning. They have high content validity (although the content validity of the PP is debatable), they are correlated with SAT scores and GPA as expected, and they are highly correlated with one another at the school level, implying that they are all measuring the same construct. Most important, all three show substantial growth from the beginning to end of college.

4 Are participants representative of their institution?

One issue that plagues all approaches to measuring student quality is the representativeness of the assessment participants. Schools often use a sampling approach; a random sample of

students is drawn, and members of the sample are asked to participate in the assessment. Sometimes schools use a census approach; all students are sent a request to participate. From a quality perspective, the issue here is not necessarily the sampling strategy, but student response to these invitations: Response rates are never 100 percent, and are usually quite low.[38]

Many scholars and practitioners believe low response rates pose a major problem for assessments, because response rates are viewed as a proxy for bias. To illustrate this line of thought, suppose the participation rate for an assessment like the CLA was high, say 70 percent. Most people would say that the data are probably representative of the student body. Suppose, however, that the participation rate was low, say 20 percent; the reasoning then is that the participants are likely to be unrepresentative of the school, and thus the resulting estimates will be biased. That is, any numbers that we would calculate, such as the percentage of students scoring proficient on an assessment, would be very different from what we might obtain if we had scores for the entire student body.

Research analyzing the relationship between response rates and bias has reached a surprising conclusion: There is no relationship.[39] In other words, bias can be quite severe with surveys with high response rates, and minimal with surveys with low response rates. What drives bias is not the response rate per se, but whether the factors that affect response are related to the survey questions.[40] Suppose a student survey on dining satisfaction yields a response rate of only 10 percent. If nonresponse is driven by factors unrelated to dining, such as students being too busy to respond, overall survey fatigue, etc., then there may be no bias whatsoever, and the survey results will perfectly mirror the opinions of the student body. On the other hand, if students dissatisfied with the cost and quality of food on campus react negatively to the invitation to participate, not wanting to help the campus office that they feel takes their money and provides little value in return, then the results would be highly skewed. Thus, any discussion of assessment participation and its effect on making quality judgments must go beyond simple response rates and focus on what drives student participation.

With this in mind, what does research on participation in the NSSE, CLA and other learn-

ing assessments tell us? First, we know that response rates vary widely across institutions. The National Survey of Student Engagement provides a useful example, because the survey instrument, sampling design, survey timing and method of administration are constant across institutions. Thus, almost all aspects of the survey process are constant, except for the makeup and culture of the student bodies.[41] Response rates among participating institutions in a given year have an extraordinary range, from as low as 14 percent to as high as 70 percent.[42]

Response rates for other assessments are not widely available. A web search found several institutional reports describing CLA administration on individual campuses; while not representative, the response rates are illustrative (see Table 2).

Table 2: CLA Response Rates

Institution	Years	First-years	Seniors	Class-based?
California State University, Pomona	2005-2006	6%	2%	No
Eastern Connecticut State University	2009-2010	75%	71%	Yes
Grand Valley State University	2005-2009	48%	29%	No
University of Missouri - St. Louis	2010-2011	23%	20%	No
University of North Carolina at Pembroke	2010-2011	11%	8%	No

As can be seen, the response rates for the CLA tend to be low, which is not surprising given the large amount of time necessary to complete the assessment. Response rates are much higher for class-based administration, in which instructors are approached and asked if they would devote class time to administering the CLA. With this approach, students are generally a captive audience, yielding a much higher response rate. The difference between the two approaches is best illustrated by the experience of Central Connecticut State University. Because it uses a class-based approach for first-years, 55 to 95 percent of students completed the CLA (response rates estimated at the class-section level). For seniors, the school took the standard approach of using several e-mail invitations, with no incentives offered. No seniors responded to the e-mail invitations.[43]

In part, the variation in response rates across schools is driven by institutional characteristics. Research indicates that response rates to college student surveys are highest for schools

that are smaller in size, are selective (high SAT scores and without a transfer mission), have large numbers of computers relative to the number of students, and are rural or located in the Midwest.[44] Individual-level research shows that survey participants tend to be female and white, have high grade-point averages, take more years of math and foreign languages in high school, and receive no financial aid. In addition to these common observables, respondents also tend to be more engaged socially during high school (belong to student groups, discuss politics, etc.), and to be more oriented toward scientific inquiry and less oriented toward status and financial success in life.[45] Many of these characteristics are likely to be correlated with student learning.

Differences between assessment participants and the overall student body are exacerbated by the sampling strategies promoted by the CLA. While their documents stress the need for a representative sample, they allow institutions to depart from random sampling and to use course sections for administering the assessment. This creates two problems. First, it is unlikely that the number of course sections used will be large enough to result in a sample representative of the population. Second, class-based administration guarantees that students who skip class will not be assessed; in other words, the primary driver of nonresponse will be attitudes toward academic achievement and learning, which are undoubtedly related to how much a student learns in college. While the CAAP and PP do not specifically mention course-based administration, this is likely taking place at some institutions, because of a lack of control over how schools create their samples and administer the instrument. This approach is quite different from that used by the NSSE, which provides institutions with detailed definitions of first-years and seniors and requires institutions to submit a population data file. NSSE staff then draw random samples for each institution, ensuring that a consistent sampling strategy is used across them all.

The exam-based instruments' lack of control over institutional sampling strategies creates another problem besides nonresponse bias. Institutions under strong accountability pressures may purposely choose student samples to maximize performance on the value-added approaches used by the exam-based instruments. This is not a minor concern. There have been several well-publicized incidents of schools submitting false data to *U.S. News & World Report*'s college rank-

ings, in an effort to improve institutional performance. One need only look at the numerous standardized testing cheating scandals in elementary and secondary education to realize that such institutional behavior is all too possible. This possibility has been dismissed by some CLA scholars, claiming: "Some critics have argued that schools may try to stack the deck, for example, by choosing their best students to take the CLA tests.... [This would not] work. To stack the deck, a school would have to find freshmen who under-perform on the CLA (relative to their SAT scores) and seniors who over-perform-a tall order at best."[46]

Such a stacking of the deck is not as difficult as it might appear. A school could administer a critical thinking test in fields where students are known to show large growth in learning, enhancing their performance. Or they could administer the test for first-years in classes, ensuring a captive audience with low motivation. For seniors, they could offer large cash inducements in an e-mail invitation to the entire senior class, ensuring a group of seniors with high motivation to excel on the test (research indicates that 5 percent of the variance in student CLA scores is accounted for by motivation[47]). Controlling for SAT in value-added models would not account for this kind of institutional behavior. An examination of the various ways that teachers and principals have altered school performance on standardized tests would suggest we should not underestimate human ingenuity and ambition in outcomes assessment.

In sum, participation rates in surveys and exam-based assessments vary widely across institutions, and individual participation is related to a variety of school and student characteristics, many of which are likely correlated with how much a student learns in college. When the factors that drive participation are correlated with student responses on surveys and exam-based measures of learning, any estimates based on those data will be biased.

5 Are relevant institutional differences taken into account?

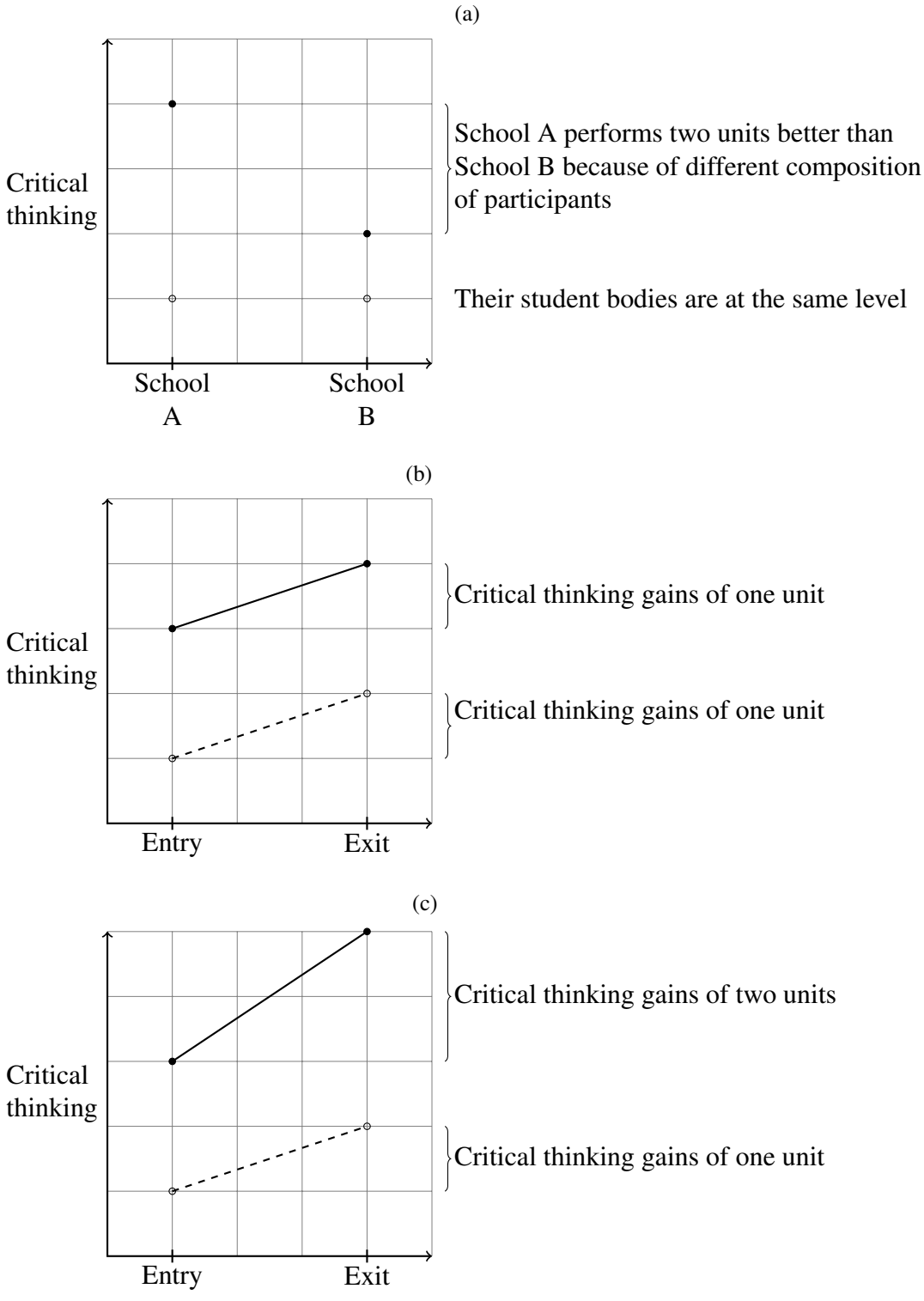
Even if we had an ideal measure of student learning that was representative of the student bodies of many institutions, the question remains whether we can meaningfully compare insti-

tutions. Harvard students will score higher than Iowa State students on a critical thinking test, because Harvard students are some of the best in the nation and began their Harvard career with excellent critical thinking and problemsolving skills. We must find some way of taking into account the different starting points of student bodies. Otherwise, more selective institutions would routinely be ranked as the highest quality simply because of the selectivity of their admissions process.

Recent research by Ernie Pascarella demonstrates the danger of not taking institutional differences into account. Using data from the Wabash national study, he and a coauthor estimate that half of the variance in the NSSE national benchmarks is explained by student and institutional characteristics.[48] Because school performance is judged by comparing individual institutions to the benchmark (i.e., the mean of all schools in the sample), this means that many schools are judged as above or below the benchmarks, not only because of what the institution is doing (fostering or inhibiting student engagement) but also because of the characteristics of the schools and their student bodies (e.g., wealth, selectivity, size, etc.). This issue has long been a concern with postsecondary researchers.[49] The danger in comparing schools with a single measurement of student learning can be seen in Panel *a* of Figure 1. The black circles show where Schools A and B score on a test of critical thinking; note that this school-level score is based on whoever is asked and then agrees to take the exam. The open circles show where each school would score if we were omniscient and knew the level of critical thinking for each member of the student body, and could then calculate the mean level for each school. The differences between the measured and actual levels of critical thinking are due to nonresponse bias: Because of their academic, demographic and attitudinal backgrounds, School A's *participants* are better critical thinkers than School B's, even if the *student bodies* as a whole are comparable. Some of this may also be due to institutional differences; School A may have offered students several hundred dollars to participate, while School B did not.

One possible solution to this problem is the estimation of a regression model with a variety of control variables at the student and school level. The idea here is that the predicted level

Figure 1: Measuring learning across institutions



of critical thinking (estimated from the model coefficients) takes into account the differences between institutions. Such an approach rests on a very strong assumption: namely, that all relevant covariates have been included in the model such that the residual captures only differences between institutions related to how institutions teach critical thinking. I am skeptical that currently available higher education datasets have the required covariates to make this assumption plausible. For example, a large body of research demonstrates that personality attributes such as conscientiousness and motivation are correlated with academic achievement, and it is difficult to believe that student bodies at institutions across the country do not vary in terms of these variables.

Another argument against this approach is the causal inference revolution currently occurring in the social sciences and education. It is partly based on the recognition that this assumption is usually untenable; studies comparing experimental estimates to regression-based estimates of interventions have found large differences, presumably due to violation of this assumption.

Methods that measure student learning at both entry and exit offer a more plausible path to comparing institutions. Several types of value-added models using more than a single cross-section have been proposed to address differences between institutions; they differ in the data and covariates used to estimate value added. The simplest approach is to administer a critical thinking exam to first-years and seniors during the same time period and then calculate the difference between the two group means. For example, a critical thinking exam administered in spring 2011 would use first-year students who entered an institution in fall 2010, as well as students who were classified as seniors in spring 2011 (meaning that they began college several years earlier).

The drawback to the concurrent data approach is that the two groups are not comparable. Seniors consist of students who have succeeded in college, at least to the extent that they were not suspended because of low academic performance, nor have they dropped out because of emotional or financial reasons. Any comparison of the two groups will overstate gains in critical thinking, because the seniors will consist of students most likely to score high on any critical thinking measure. Research using critical thinking tests and concurrent samples of first-years and seniors consistently finds that a) seniors score higher and b) seniors have higher SAT scores than

first-years. I note that because of attrition, seniors undoubtedly differ from first-years on many other variables besides SAT scores, particularly unobservables such as ambition, conscientiousness and academic motivation.

A second approach, which I term *concurrent value-added*, uses concurrent data and institution-level means for SAT scores and critical thinking tests; this approach is currently used by the Voluntary System of Accountability as well as the CAAP, CLA and PP. Two regression models are estimated, one for first-years and one for seniors, predicting CLA scores using only SAT scores as a covariate.[50] The difference between the residuals from each model is used to estimate an institution's value added. In essence, this approach tries to overcome the objection raised previously: that seniors are different from first-years, and that we can't simply compare their mean scores.

One way to think about these residuals is that they estimate how different an institution's mean CLA score is from the mean CLA score of other institutions with the exact same SAT score. So we can determine, for example, how far above (or below) first-years at a school are compared with first-years at a similar institution. We can make the same determination for seniors. Next, we can compare how much the institution's first-years and seniors differ on these differences. If seniors show a larger difference than first-years, then this must be due to a larger growth in critical thinking skills among students at that institution.

This approach suffers from the same problem that the cross-sectional residual models described above suffer from. It assumes that the only relevant difference between first-years and seniors that needs to be modeled is standardized test scores. Yet standardized test scores are only one variable that differs between first-years and seniors. If other factors, such as personality and motivation, differ between first-years and seniors, and if they operate in such a way that their effect may differ across institutions (which is likely), then this approach will yield biased estimates of institutional value added.

Another way to think about this is to ask the following question: Why should we control only for standardized test scores? Are there not other variables that could be determining both

attrition during college and how well a student scores on a test of critical thinking? When viewed from this perspective, it becomes clear that these models are seriously underspecified, and that institutional results should be viewed with some skepticism.[51]

There is also a theoretical drawback to comparing concurrent groups of first-years and seniors. With this approach, critical thinking scores for seniors reflect events for the past four to six years, while for first-years they reflect events from the current year. Suppose that an institution did not place much emphasis on critical thinking skills until the current year, resulting in a big increase in first-year scores from what they would have been under the prior regime. This increase might be equivalent to the critical thinking gains that seniors have accrued during the past four to six years. Any analysis comparing scores for the two concurrent groups would then reveal that no learning is taking place at that institution! Proponents favor this approach over a true longitudinal design because of worries that institutions do not wish to wait four years for results. Yet many institutions use surveys of graduating seniors for internal accountability purposes, asking seniors to rate their college experience for the previous four to six years in many different areas. How this is any different from analyzing the results of a true longitudinal design is never explained.

The third approach, which I term *longitudinal value added*, also measures critical thinking for first-years and seniors but measures the same student at entry and exit. The advantages of this approach can be seen in Panel *b* of Figure 1. The two lines represent two students who differ in critical thinking at entry because of hundreds, perhaps even thousands, of variables. If these variables, and their effects, are stable during college, then by comparing their scores at exit to those at entry, we are in effect controlling for these unmeasured variables. So while the students differ in background characteristics, they show the exact same growth in critical thinking. This demonstrates the main advantage of a true longitudinal design: We can worry less about differences in student background across institutions, because each student in effect serves as her own control.

We can also think of the two lines in Panel *b* as showing the mean scores of a single institution with assessment participants (solid circles and line) who are quite different from the main

student body (open circles and dotted line) because of nonresponse bias. As long as the factors driving nonresponse are similar at entry and exit, true longitudinal designs will also help take into account nonresponse bias across institutions when estimating learning gains: The gains are the same for both groups.

Longitudinal designs do rely on the assumption that growth in learning is the same for assessment participants and nonparticipants. Panel *c* illustrates learning gains at a single institution, with the dashed line representing the entire student body and the solid line assessment participants. If students recruited to participate in a longitudinal study at entry differ from the student body in how much they can gain, then learning growth estimates for a school will be biased. For example, suppose that participants are overwhelmingly female, white, not on financial aid, high-GPA students, oriented toward scientific inquiry, whereas the general student body is on average much lower on these characteristics, as the survey methods literature suggests would happen. As long as we expect learning growth rates for females, whites, high academic ability, high socioeconomic status and investigative students to be the same as for males, nonwhites, low academic ability, low socioeconomic status students oriented toward business, then nonresponse should not pose a problem. But the assumption that the learning growth rates are similar for these students is implausible; Arum and Roksa find CLA growth rates to vary by parental education, race and financial aid.

There have been some limited efforts to measure learning over time.[52] The main objection, besides timing, is expense. If an institution has a 70 percent graduation rate and needs complete longitudinal data at exit for 100 students, then it would need to test 143 students at entry. While cost is certainly a factor in assessment decisions, the accuracy of any assessment must also play a role. Given the much stronger methodological basis of longitudinal designs, and the mistaken concerns about timing, this approach is worth the added costs.

Finally, any approach using value-added scores must take into account that these estimates contain uncertainty.[53] The concurrent value-added approach simply estimates the difference between expected growth for first-years and expected growth for seniors, and then ranks

schools by how large this difference is. But these are estimates based on samples, and so there is some uncertainty as to what the exact amount of the difference is. Typically this uncertainty is estimated with a confidence interval, most commonly seen in public opinion polls. Support for a specific policy may be reported at 60 percent, plus or minus 4 percent. Although news reports tend to focus on the point estimate, all we can really say in this situation is that support for that policy among the population lies between 56 percent and 64 percent.[54]

Taking into account this uncertainty is important for two reasons. First, results from the exam-based instruments are often expressed as a number representing the amount of value added. If the number is greater than zero, then the school is seen as adding value: that is, its students have increased their critical thinking skills compared with what growth should have been, taking into account standardized test scores.[55] But this estimated difference could have easily arisen because of random chance, meaning that the school actually does not add value (in statistical parlance, the school's confidence interval brackets zero). The same issue arises when schools are compared against a benchmark estimated from a national sample (like the NSSE) or a set of performance standards.[56] Second, any ranking of schools based on these assessments is based on the premise that a small increase in value added means one school should be ranked above another school, even if their confidence intervals overlap, indicating that their estimated value-added scores are statistically indistinguishable.

A recent HLM analysis of CLA data estimated value-added scores and confidence intervals for individual schools. About two-thirds of the confidence intervals bracketed zero, which in this case meant that schools were performing as expected. Yet the point estimates of most of these schools were quite a bit above or below zero.

In sum, for any measure of student learning to be comparable across institutions, institutional differences must be taken into account. Simple cross-sectional models cannot accomplish this, and the concurrent value-added approach of comparing first-years and seniors at the same point in time is also problematic. Most promising are true longitudinal designs, in which the same student's critical thinking is measured at both entry and exit. Any school-level estimates

of learning growth must take into account uncertainty created by sampling and estimation procedures, such that schools are judged based on confidence intervals rather than misleading point estimates.

6 Recommendations

6.1 For institutional policymakers

Recommendation 1: Discontinue use of the NSSE and other college student surveys to assess learning

Given the lack of theoretical and empirical evidence that the NSSE and other surveys are measuring behaviors related to learning or student learning gains, it makes little sense to continue the use of these surveys for institutional assessments of learning. Even if a valid measure of student engagement could be created, it is not clear how to take into account the differences among student bodies across institutions. It is not possible to measure student engagement at entry, as students must be actively participating in college in order to measure their engagement. The cross-sectional approach of the NSSE makes it difficult to disentangle the effects of student inputs from actual student learning on their institutional benchmarks. Engagement proponents have not addressed this issue, and it seems as if the only possibility would be a cross-sectional type of value-added model. These models rely on assumptions that can rarely be met in practice, which is one reason the K-12 sector has abandoned cross-sectional value-added models in favor of models that use two or more observations of student achievement over time.

The lack of validity evidence for these surveys raises the question of why they are so popular with institutions. The answer is that they are relatively cheap to administer on a per-student basis. This leads to a corollary to Recommendation 1: *Institutions should spend far more money on assessing student learning, and governance agencies should push them to do so.* It is surprising that institutions with multimillion-dollar budgets can seem to scrape together only a few thousand dollars to spend on assessing what should be the central activity of every college: student

learning. Given the costs of the exam-based assessments of critical thinking, and the necessity of large cash payments to motivate students to participate, institutions must be willing to expend much more resources in this critical area. Spending a few dollars per student to measure the chief outcome of college, critical thinking, makes little sense in the current age of accountability.

Recommendation 2: Use exam-based instruments to assess learning

The validity evidence for all three exam-based instruments is remarkably similar; they are correlated with SAT and GPA as expected, are highly correlated with one another, and exhibit large growth from the beginning to end of college. Of these, the CAAP is probably the most useful, for the following reasons.

First, the Test Validity Study demonstrates that school-level correlations between the CAAP, CLA and PP are high, suggesting that these instruments are measuring the same construct. These high correlations then raise the question of why the CLA approach is necessarily better. The CLA originally attracted attention because its tasks somehow seemed more authentic than the multiple-choice testing approach of the CAAP and PP. Reliance on face validity is problematic, as social scientists well know, because at its essence it relies more on gut feelings than any quantifiable set of measures. If we accept the face validity of the CLA because of its unique use of multiple, related artifacts, the high correlations demonstrated by the Test Validity Study indicate that such lengthy tasks involving multiple artifacts may be unnecessary, and the same results can be achieved with multiple-choice tests of critical thinking.

Second, the CAAP measures only one construct, while the CLA measures both critical thinking and writing ability. Whether this really matters depends largely on how the results are used within an institution. If institutions want to use results to guide programmatic change, it may be difficult to determine exactly what needs to be changed given a poor performance on the CLA. On the other hand, if governance bodies are simply seeking an overall performance measure for institutions, then the combination of the two constructs may not be a drawback, given that everyone would agree students should leave college as better writers as well as better critical thinkers. Given the limited nature of the PP, its content validity is questionable. Faculty at an in-

stitution reviewing the instrument would likely conclude that it does a poor job measuring critical thinking, especially when compared with the readings used by the CAAP and CLA.

Third, the CAAP is more reliable than the CLA at the student level and has similar reliabilities at the school levels. Proponents of the CLA argue that this is to be expected, given that the CLA measures students with only one item, rather than multiple items. While true, given a choice between two instruments, most researchers would choose the instrument that is reliable at both the student and school levels.

Fourth, multiple-choice exams take less time and are less of a burden for students, which makes recruitment easier. This in turn should increase participation rates, which might make participant samples more representative. The CLA Performance Task requires 90 minutes, while the critical thinking portion of the CAAP requires less than half that time, 40 minutes.

Fifth, the CAAP is less expensive for schools, an important consideration given the current budgetary situation and historically low amount of resources dedicated to assessment by institutions. The CLA currently charges $6,600 to administer its instrument twice, to two samples of 100 students. The cost for a similar administration of the CAAP would be $2,800, less than half the cost of the CLA. Given the need for large samples of students to obtain more precise estimates of institutional performance (i.e., narrower confidence intervals), and the few resources generally available for assessment within institutions, the lower cost of the CAAP is an important consideration when evaluating instruments.

Sixth, multiple-choice testing allows for the measurement of motivation in terms of time spent on test items. We cannot forget that these instruments are low-stakes exams, and that there is no conclusive evidence that test motivation remains the same at college entry and exit. Computer administration of a critical thinking test allows for response-time motivation filtering and the use of non-attitudinal motivation tests such as speed coding, offering ways to measure and take into account differing levels of test motivation across time and schools.[57]

Recommendation 3: Adopt true longitudinal designs to measure student learning

In order to make reasonably accurate quality judgments about institutions, we must at

a minimum take into account the heterogeneity of student bodies, as well as differential nonresponse bias among institutions. While we might try to correct for these differences using regression models with nonresponse weights, such an approach can correct only for observable student differences. However, student bodies and assessment participants vary in many ways, such as personality and motivation, that will also affect how much they learn during college. Rather than attempt to measure all of these variables, a simpler approach would be to measure critical thinking for the same set of students at entry and exit, and then use the difference over time as a measure of student learning. Such difference-in-difference estimators control for myriad student unobservables that vary across institutions.

This approach would yield more buy-in from faculty and other stakeholders, because the longitudinal design is easily explained and fits with many people's intuitive beliefs about learning. Schools ranked low in quality by a complex multivariate model with survey weights would probably reject the results, arguing that a different set of independent variables and weights would yield more accurate results. Instead of engaging schools with a discussion about teaching and learning at their institutions, we would likely be drawn into a bitter debate over methods. Although a true longitudinal design is costlier than a concurrent data approach, benefits from controlling for unobservables and the transparency of the method far outweigh the additional monetary costs.

One drawback to this approach is student attrition; some schools lose significant numbers of students by their senior year. Surveys administered during the senior year, such as the NSSE and HERI College Senior Survey, also suffer from this problem, so from a comparative perspective it does not appear that measuring students at entry and exit and taking the difference is any worse than current approaches. More problematic is what this attrition means for comparative purposes. One could imagine a school performing worse on a measure because it does a good job retaining at-risk students, while a similar school that spends little effort on retaining at-risk students would show better performance, because its at-risk students would not appear in its pool of participants.

Recommendation 4: Consider alternative ways to turn low-stakes exams into high-stakes exams

Student motivation to excel while participating in these measures is undoubtedly low. The instruments take 40–90 minutes to complete, and the CLA requires students to comprehend and synthesize several different readings. Thus, it is an open question whether students are fully grappling with the material, or doing just enough to get by. Possibilities here include substantial monetary rewards based on performance, university honors based on performance, or even the use of these measures as a graduation exit exam.

Finally, both survey and exam-based measures of learning suffer from nonresponse bias: Only certain types of students are likely to participate in these assessments. This will become more problematic in the future, as available evidence suggests that survey response rates continue to fall as the availability of web survey software increases survey fatigue. Similar to the issue of student performance on exam-based measures, the problem here is one of student motivation. Many schools are already offering financial compensation for participation, so one possible solution is to require participation as part of the academic experience. Some schools are already doing so and listing required participation in their university catalogs. As with offering university honors based on exam performance, this would require a sea change on many campuses, in terms of how we think about student evaluation and assessment.

While we can debate many of the issues around current efforts to measure student learning, in the end we face the fundamental problem of motivation: How can we motivate a representative group of students to participate in these assessments, and then exert their maximum cognitive effort while participating?

Recommendation 5: Institutions and system leaders should pressure assessment producers to standardize institutional sampling and student incentives, in order to provide more comparable data across institutions

By taking the sampling process out of the hands of school administrators, the NSSE prevents schools from targeting specific groups of students to boost institutional performance. Such

behavior by institutions is a real concern; we should not underestimate the pressure that college administrators face for better performance on institutional metrics.[58] Inter-institutional benchmarks and value-added models require comparable data, and such data are suspect when institutions can not only draw their own samples but even direct invitations to participate to specific course sections of students. Given research indicating that motivation affects performance on the CLA, and that students prefer some types of incentives over others, some consistent framework of participation rewards must be established (I note that these issues are not limited to the CLA but apply to other measures of critical thinking).

6.2 For research funders

Recommendation 6: Fund research in three areas: general validation research, the effects of motivation on nonresponse and test performance, and the use of value-added models in higher education

First, as can be seen in this review, there is still relatively little evidence that exam-based measures of critical thinking are actually measuring critical thinking. Far more studies are needed to assure policymakers and stakeholders that institutions should, in essence, stake everything on these assessments. For example, almost no work has been done that demonstrates a causal link between student experiences during their college career and their growth in critical thinking, besides the estimation of simple correlations with GPA and research showing differences across major fields. It is also not clear how well these tests perform for schools with students who begin school with a very high (or low) level of critical thinking skills; in other words, can they be considered valid for the full spectrum of students?

Second, we know that motivation and other student characteristics likely play a role in both the decision to participate and the amount of cognitive effort exerted during assessments. Yet we know almost nothing about how participants and nonparticipants differ on attitudinal constructs, and how much of an effect these differences have on institutional estimates of critical thinking (one exception is the study by Swerdzewski and colleagues, which found substantial

differences between test-takers and test-avoiders).[59] We also know little about what types of incentives best motivate students.

Third, the value-added models proposed by the CLA and others are simplistic in nature and lack strong statistical and empirical work supporting their use in evaluating postsecondary institutions. This is in stark contrast to research on value added in the K-12 sector, which has established a large body of work supporting the use of value added models to measure value added by teachers and schools.[60] Far more research is needed in this area before we can begin to use these models to evaluate and rank postsecondary institutions in terms of their quality. The effect of attrition on value-added measures, for example, has been relatively unexplored. It would also be useful to explore ways in which the reliability of difference estimators could be increased, either through combining several years of data or through alternative methodological approaches.[61]

Recommendation 7: Fund researchers who are working outside of organizations producing assessments of learning

Almost all of the validity research on the self-reported behaviors, self-reported learning gains and direct measures of student learning discussed in this review have been conducted by researchers heavily involved with the organizations that are designing, marketing and administering these instruments. These relationships pose a clear conflict of interest. Given the myriad number of choices that must be made with any empirical analysis, these conflicts of interest could be consciously or unconsciously affecting choices made by these researchers. Unlike the field of medicine, which has openly struggled with issues surrounding research funded by drug companies and doctors recommending procedures using medical devices created by their own companies, the field of postsecondary research has largely ignored this topic.

Given the large sums of money at stake (e.g., the gross revenue from the 2011 administration of the NSSE was approximately $3.5 million[62]), research by independent scholars is essential if the validity evidence of these assessments is to gain credibility among institutions and educational stakeholders.

Notes

1. Richard Arum and Josipa Roksa, *Academically Adrift: Limited Learning on College Campuses* (University of Chicago Press, 2011).

2. Ibid.

3. M. Scriven and R.W. Paul, *Critical Thinking as Defined by the National Council for Excellence in Critical Thinking*, Presented at the 8th Annual International Conference on Critical Thinking and Education Reform, Summer, 1987.

4. I focus on these surveys because they are by far the most popular surveys used by colleges and universities in the U.S. In 2011, 751 institutions administered the NSSE, 435 the CCSSE, 58 the HERI Your First College Year Survey, and 89 the HERI College Senior Survey.

5. Stephen R. Porter, "Do college student surveys have any validity?" *Review of Higher Education* 35 (2011): 45–76.

6. Examples taken from the latest versions of the NSSE, CCSSE and the HERI College Senior and Your First College Year surveys.

7. N.M. Bradburn, Lance J. Rips, and S.K. Shevell, "Answering autobiographical questions: The impact of memory and inference on surveys," *Science* 123 (1987): 157–161; Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski, *The Psychology of Survey Response* (Cambridge University Press, 2000).

8. See, e.g., the NSSE and CCSSE; HERI surveys using these questions include the CIRP Freshman, Your First College Year, and College Senior surveys.

9. Wendy E. Pentland et al., eds., *Time Use Research in the Social Sciences* (Kluwer, 2002); John P. Robinson, *How Americans use time: a social-psychological analysis of everyday behavior* (Praeger, 1977); John P. Robinson, "The Validity and Reliability of Diaries versus Alternative Time Use Measures," in *Time, Goods and Well-Being*, ed. F. Thomas Juster and Frank Stafford (University of Michigan Press, 1985), 33–62; John P. Robinson and Ann Bostrom, "The overestimated workweek: What time diary measures suggest," *Monthly Labor Review* 11 (1994): 11–23; Ralph Stinebrickner and Todd R. Stinebrickner, "Time use and college outcomes," *Journal of Econometrics* 121 (2004): 243–269.

10. G. Bishop, R. Oldendick, and A. Tuchfarber, "Opinions on fictitious issues: The pressure to answer survey questions," *Public Opinion Quarterly* 50 (1986): 240–250; H. Schuman and S. Presser, *Questions*

and answers in attitude surveys: Experiments in question form, wording, and context (Academic Press, 1981); Patrick Sturgis and Patten Smith, "Fictitious issues revisited: Political interest, knowledge and the generation of nonattitudes," *Political Studies* 58 (2010): 66–84.

11. Porter, "Do college student surveys have any validity?"

12. Stephen R. Porter and Andrew A. Ryder, *What is the appropriate time frame for measuring the frequency of educational activities?* Paper presented at European Survey Research Association annual meeting, Lausanne, Switzerland, 2011.

13. Roger Tourangeau, Mick P. Couper, and Frederick Conrad, "Spacing, position, and order: Interpretive heuristics for visual features of survey questions," *Public Opinion Quarterly* 68 (2004): 368–393; Roger Tourangeau, Mick P. Couper, and Frederick Conrad, "Color, labels, and interpretive heuristics for response scales," *Public Opinion Quarterly* 71 (2007): 91–112.

14. Porter, "Do college student surveys have any validity?"; Nicholas A. Bowman and Patrick L. Hill, "Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains," *New Directions for Institutional Research* 150 (2011): 73–85; but for a contrary view see R.M. Gonyea and Angie Miller, "Clearing the AIR about the use of self-reported gains in institutional research," *New Directions for Institutional Research* 150 (2011): 99–111.

15. Robert M. Carini, George D. Kuh, and Stephen P. Klein, "Student engagement and student learning: Testing the linkages," *Research in Higher Education* 47 (2006): 1–32; Thomas F. Nelson Laird et al., *The predictive validity of a measure of deep approaches to learning*, paper presented at the Association for the Study of Higher Education annual meeting, Jacksonville, Fla. 2008; Ernest T. Pascarella and Tricia A. Seifert, *Validation of the NSSE benchmarks and deep approaches to learning against liberal arts outcomes*, Paper presented at the Association for the Study of Higher Education annual meeting, Jacksonville, FL. 2008.

16. Corbin M. Campbell and Alberto F. Cabrera, "How sound is NSSE?: Investigating the psychometric properties of NSSE at a public, research-extensive institution," *Review of Higher Education* 35 (2011): 77–103; M.B. Fuller, M.A. Wilson, and R.M. Tobin, "The national survey of student engagement as a predictor of undergraduate GPA: a cross-sectional and longitudinal examination," *Assessment & Evaluation in Higher Education* 36 (2011): 735–748; Jonathan Gordon, Joe Ludlum, and J. Joseph Hoey, "Validating NSSE Against Student Outcomes: Are They Related?" *Research in Higher Education* 49 (2008): 19–39;

Amy M. Korzekwa, "An examination of the predictive validity of National Survey of Student Engagement benchmarks and scalelets" (PhD diss., University of New Mexico, 2010); George D. Kuh et al., "Unmasking the effects of student engagement on first-year college grades and persistence," *Journal of Higher Education* 79 (2008): 540–563; National Survey of Student Engagement, *Validity: Predicting Retention and Degree Progress*, technical report (Center for Postsecondary Research, Indiana University, 2010).

17. Tourangeau, Rips, and Rasinski, *The Psychology of Survey Response*.

18. Vague quantifier refers to quantities that are not clearly defined, and whose meaning varies across respondents. Typical response scales for these questions are *very much, quite a bit, some, very little* (NSSE surveys) and *much stronger, stronger, no change, weaker, much weaker* (HERI surveys).

19. Nicholas A. Bowman, "Can 1st-year students accurately report their learning and development?" *American Educational Research Journal* 47 (2010): 466–496; Nicholas A. Bowman, "Assessing learning and development among diverse college students," *New Directions for Institutional Research* 145 (2010): 53–71; Nicholas A. Bowman, "Validity of college self-reported gains at diverse institutions," *Educational Researcher* 40 (2011): 22–24.

20. Descriptions of the instruments are taken from the Test Validity Study and sample questions posted on the ACT, CAE and ETS websites.

21. While the CLA provides subscores for each student in four areas (Analytic Reasoning & Evaluation, Writing Effectiveness, Writing Mechanics, and Problem Solving; see http://www.collegiatelearningassessment.org/files/CLAScoringCriteria.pdf), it is still unclear whether writing ability can be completely disentangled from critical thinking ability.

22. One can also argue that all critical thinking instruments to some extent measure reading skills as well.

23. Roger Benjamin and Marc Chun, "A new field of dreams: The Collegiate Learning Assessment project," *Peer Review* (2003): 26–29; Collegiate Learning Assessment, *Frequently Asked Technical Questions 2007-2008*, technical report (Council for Aid to Education, 2008); Stephen P. Klein et al., "An approach to measuring cognitive outcomes across higher education institutions," *Research in Higher Education* 46 (2005): 251–276; Stephen P. Klein et al., "The Collegiate Learning Assessment: Facts and Fantasies," *Evaluation Review* 31 (2007): 415–439.

24. Ou Lydia Liu, "Measuring value-added in higher education: conditions and caveats - results from using the Measure of Academic Proficiency and Progress (MAPP)," *Assessment & Evaluation in Higher*

Education 36 (2011): 81–94; Diana Marr, *Validity of the Academic Profile*, technical report (Educational Testing Service, 1995).

25. Marr, *Validity of the Academic Profile*.

26. Arum and Roksa, *Academically Adrift: Limited Learning on College Campuses*.

27. ACT, *Collegiate Assessment of Academic Proficiency Technical Report 2007-2008*, technical report (ACT, 2008).

28. Stephen P. Klein, Ou Lydia Liu, and James Sconing, *Test Validity Study Report*, technical report (Council for Aid to Education, 2009).

29. I say *suggest* rather than *demonstrate* because these high correlations are consistent with alternative explanations: "... a high correlation between two tests is necessary but not sufficient for concluding that the tests measure the same thing and that their scores can legitimately be compared. After all, it may simply be that schools with students possessing certain skills (e.g., scientific reasoning) also tend to be schools with students possessing other skills (e.g., reading comprehension)." (Jeffrey T. Steedle, Heather Kugelmass, and Alex Nemeth, "What do they measure? Comparing three learning outcomes assessments," *Change* (2010): 33–37, p. 36).

30. Jeffrey T. Steedle, *Improving the reliability and interpretability of value-added scores for postsecondary institutional assessment programs*, Paper presented at the American Educational Research Association annual meeting, Denver, CO, 2010.

31. Arum and Roksa, *Academically Adrift: Limited Learning on College Campuses*; Stephen P. Klein, *The Lumina Longitudinal Study: Summary of procedures and findings*, technical report (Council for Aid to Education, 2009).

32. Ernest T. Pascarella et al., "How robust are the findings of Academically Adrift?" *Change* (2011): 20–24.

33. Klein, Liu, and Sconing, *Test Validity Study Report*.

34. Klein, *The Lumina Longitudinal Study: Summary of procedures and findings*.

35. Liu, "Measuring value-added in higher education: conditions and caveats - results from using the Measure of Academic Proficiency and Progress (MAPP)."

36. Carolyn J. Hill et al., *Empirical Benchmarks for Interpreting Effect Sizes in Research*, technical report (MDRC, 2007).

37. Angela Lee Duckworth et al., "Role of test motivation in intelligence testing," *Proceedings of the National Academy of Sciences* (in press): 1–5; Hanna Eklof, "Skill and will: Test-taking motivation and assessment quality," *Assessment in Education: Principles, Policy & Practice* 17 (2010): 345–356; Steven L. Wise and Christine E. DeMars, "Examinee noneffort and the validity of program assessment results," *Educational Assessment* 15 (2010): 27–41.

38. There is a technical issue with these two approaches related to quality: Larger sample sizes yield more precise estimates of learning at an institution, which can make a difference when calculating value added.

39. Robert M. Groves, "Nonresponse rates and nonresponse bias in household surveys," *Public Opinion Quarterly* 70 (2006): 646–675; Robert M. Groves and Emilia Peytcheva, "The impact of nonresponse rates on nonresponse bias: A meta-analysis," *Public Opinion Quarterly* 72 (2008): 167–189.

40. Groves, "Nonresponse rates and nonresponse bias in household surveys."

41. The one exception is the use of publicity and incentives such as lottery prizes to increase response rates. These are used by some but probably not all institutions. These efforts likely have little effect on response rates. There is no documented evidence that publicity increases response rates, and rigorous experimental designs testing the effect of prizes and low-value post-paid incentives on one-shot surveys of random samples find small to no effects on response rates. See Daniel R. Petrolia and Sanjoy Bhattacharje, "Revisiting incentive effects: Evidence from a random-sample mail survey on consumer preferences for fuel ethanol," *Public Opinion Quarterly* 73 (2009): 537–550; Stephen R. Porter and Michael E. Whitcomb, "The impact of lottery incentives on student survey response rates," *Research in Higher Education* 44 (2003): 389–407.

42. Stephen R. Porter and Paul D. Umbach, "Student survey response rates across institutions: Why do they vary?" *Research in Higher Education* 47 (2006): 229–247.

43. Braden J. Hosch, *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 meeting, Chicago, IL, 2010.

44. Matt Jans and Anthony Roman, "National response rates for surveys of college students: Institutional, regional, and design factors," in *Proceedings of the Joint Statistical Meetings of the American Statistical Association, Section on Survey Research Methods* (2007); Porter and Umbach, "Student survey response rates across institutions: Why do they vary?"

45. Eric L. Dey, "Working with low survey response rates: The efficacy of weighting adjustments," *Re-*

search in Higher Education 38 (1997): 215–227; Stephen R. Porter and Michael E. Whitcomb, "Non-response in student surveys: the role of demographics, engagement and personality," *Research in Higher Education* 46 (2005): 127–152; Linda J. Sax, Shannon K. Gilmartin, and Alyssa N. Bryant, "Assessing response rates and nonresponse bias in web and paper surveys," *Research in Higher Education* 44 (2003): 409–432.

46. Stephen P. Klein et al., "Assessing school effectiveness," *Evaluation Review* 32 (2008): 114.

47. Jeffrey T. Steedle, *Incentives, motivation, and performance on a low-stakes test of college learning*, Paper presented at the American Educational Research Association annual conference, Denver, CO, 2010.

48. Ernest T. Pascarella and Ryan Padgett, *Using institution-level NSSE benchmarks to assess engagement in good practices: A cautionary note*, manuscript, University of Iowa, 2011. See also Alexander W. Astin and Jenny J. Lee, "How risky are one-shot cross-sectional assessments of undergraduate students?" *Research in Higher Education* 44, no. 6 (2003): 657–672.

49. Ibid.

50. ACT scores are converted using a concordance table; I use SAT as shorthand for SAT and ACT scores.

51. Two recent studies use HLM and student-level data to estimate concurrent value added. Interestingly, the CLA paper uses only standardized test scores (both individual-level and school-level) as covariates, while the PP paper includes admission rates and the presence of graduate programs. See Ou Lydia Liu, "Value-added assessment in higher education: A comparison of two methods," *Higher Education* 61 (2011): 445–461; Steedle, *Improving the reliability and interpretability of value-added scores for postsecondary institutional assessment programs*. Neither paper theoretically justifies their model specification, and it is likely that the two models would yield a different ranking of institutions, because of the differing sets of independent variables. See Stephen R. Porter, "The robustness of the graduation rate performance indicator used in the U.S. News & World Report college rankings," *International Journal of Educational Advancement* 1 (2000): 145–164 for more problems with the cross-sectional approach to value added.

52. Klein, *The Lumina Longitudinal Study: Summary of procedures and findings*; Pascarella et al., "How robust are the findings of Academically Adrift?"

53. Steedle, *Improving the reliability and interpretability of value-added scores for postsecondary institutional assessment programs*.

54. Specifically, we say that we are 95 percent confident that support lies between 56 percent and 64

percent. Ninety-five percent confidence means that if we sampled the population over and over again, and estimated a confidence interval for each sample, our interval would bracket actual support in the population in 95 percent of the samples.

55. Interpretation of value-added scores can be tricky and depends on how they have been estimated. This interpretation is how we would view simple difference scores as measures of value added. Interpretation of a positive number using the CLA residual approach is different; a positive number indicates that students learned more than expected (that is, more than the mean learning for all schools).

56. Chaitra M. Hardison and Anna-Marie Vilamovska, *The Collegiate Learning Assessment: Setting Standards for Performance at a College or University*, technical report (RAND Corporation, 2009).

57. Carmit Segal, *Working when no one is watching: Motivation, test scores, and economic success*, unpublished manuscript, Universitat Pompeu Fabra, 2011; Peter J. Swerdzewski, J. Christine Harmes, and Sara J. Finney, "Two approaches for identifying low-motivated students in a low-stakes assessment context," *Applied Measurement in Education* 24 (2011): 162–188; Wise and DeMars, "Examinee noneffort and the validity of program assessment results."

58. For example, the provost at Iona College was recently discovered to have submitted false data to U.S. News, the federal government and bond rating agencies. Not surprisingly, most of the false data elements were those used in college rankings. Doug Lederman, "Disingenuous Data," *Inside Higher Ed* (Nov. 9, 2011).

59. Peter J. Swerdzewski, J. Christine Harmes, and Sara J. Finney, "Skipping the Test: Using Empirical Evidence to Inform Policy Related to Students Who Avoid Taking Low-Stakes Assessments in College," *Journal of General Education* 58 (2009): 167–195.

60. Douglas N. Harris, *Value-Added Measures in Education* (Harvard Education Press, 2011).

61. Jeffrey T. Steedle, "Selecting value-added models for postsecondary institutional assessment," *Assessment & Evaluation in Higher Education* (in press).

62. Calculation by the author based on the number of 2011 participating institutions and fee structure posted on the NSSE website.

References

ACT. *Collegiate Assessment of Academic Proficiency Technical Report 2007-2008*. Technical report. ACT, 2008.

Arum, Richard, and Josipa Roksa. *Academically Adrift: Limited Learning on College Campuses*. University of Chicago Press, 2011.

Assessment, Collegiate Learning. *Frequently Asked Technical Questions 2007-2008*. Technical report. Council for Aid to Education, 2008.

Astin, Alexander W., and Jenny J. Lee. "How risky are one-shot cross-sectional assessments of undergraduate students?" *Research in Higher Education* 44, no. 6 (2003): 657–672.

Benjamin, Roger, and Marc Chun. "A new field of dreams: The Collegiate Learning Assessment project." *Peer Review* (2003): 26–29.

Bishop, G., R. Oldendick, and A. Tuchfarber. "Opinions on fictitious issues: The pressure to answer survey questions." *Public Opinion Quarterly* 50 (1986): 240–250.

Bowman, Nicholas A. "Assessing learning and development among diverse college students." *New Directions for Institutional Research* 145 (2010): 53–71.

———. "Can 1st-year students accurately report their learning and development?" *American Educational Research Journal* 47 (2010): 466–496.

———. "Validity of college self-reported gains at diverse institutions." *Educational Researcher* 40 (2011): 22–24.

Bowman, Nicholas A., and Patrick L. Hill. "Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains." *New Directions for Institutional Research* 150 (2011): 73–85.

Bradburn, N.M., Lance J. Rips, and S.K. Shevell. "Answering autobiographical questions: The impact of memory and inference on surveys." *Science* 123 (1987): 157–161.

Campbell, Corbin M., and Alberto F. Cabrera. "How sound is NSSE?: Investigating the psychometric properties of NSSE at a public, research-extensive institution." *Review of Higher Education* 35 (2011): 77–103.

Carini, Robert M., George D. Kuh, and Stephen P. Klein. "Student engagement and student learning: Testing the linkages." *Research in Higher Education* 47 (2006): 1–32.

Dey, Eric L. "Working with low survey response rates: The efficacy of weighting adjustments." *Research in Higher Education* 38 (1997): 215–227.

Duckworth, Angela Lee, P.D. Quinn, D.R. Lynam, R. Loeber, and M. Stouthamer-Loeber. "Role of test motivation in intelligence testing." *Proceedings of the National Academy of Sciences* (in press): 1–5.

Eklof, Hanna. "Skill and will: Test-taking motivation and assessment quality." *Assessment in Education: Principles, Policy & Practice* 17 (2010): 345–356.

Fuller, M.B., M.A. Wilson, and R.M. Tobin. "The national survey of student engagement as a predictor of undergraduate GPA: a cross-sectional and longitudinal examination." *Assessment & Evaluation in Higher Education* 36 (2011): 735–748.

Gonyea, R.M., and Angie Miller. "Clearing the AIR about the use of self-reported gains in institutional research." *New Directions for Institutional Research* 150 (2011): 99–111.

Gordon, Jonathan, Joe Ludlum, and J. Joseph Hoey. "Validating NSSE Against Student Outcomes: Are They Related?" *Research in Higher Education* 49 (2008): 19–39.

Groves, Robert M. "Nonresponse rates and nonresponse bias in household surveys." *Public Opinion Quarterly* 70 (2006): 646–675.

Groves, Robert M., and Emilia Peytcheva. "The impact of nonresponse rates on nonresponse bias: A meta-analysis." *Public Opinion Quarterly* 72 (2008): 167–189.

Hardison, Chaitra M., and Anna-Marie Vilamovska. *The Collegiate Learning Assessment: Setting Standards for Performance at a College or University*. Technical report. RAND Corporation, 2009.

Harris, Douglas N. *Value-Added Measures in Education*. Harvard Education Press, 2011.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark Lipsey. *Empirical Benchmarks for Interpreting Effect Sizes in Research*. Technical report. MDRC, 2007.

Hosch, Braden J. *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 meeting, Chicago, IL, 2010.

Jans, Matt, and Anthony Roman. "National response rates for surveys of college students: Institutional, regional, and design factors." In *Proceedings of the Joint Statistical Meetings of the American Statistical Association, Section on Survey Research Methods*. 2007.

Klein, Stephen P. *The Lumina Longitudinal Study: Summary of procedures and findings*. Technical report. Council for Aid to Education, 2009.

Klein, Stephen P., Ou Lydia Liu, and James Sconing. *Test Validity Study Report*. Technical report. Council for Aid to Education, 2009.

Klein, Stephen P., George D. Kuh, Marc Chun, Laura Hamilton, and Richard Shavelson. "An approach to measuring cognitive outcomes across higher education institutions." *Research in Higher Education* 46 (2005): 251–276.

Klein, Stephen P., David Freedman, Richard Shavelson, and Roger Bolus. "Assessing school effectiveness." *Evaluation Review* 32 (2008): 114.

Klein, Stephen P., Roger Benjamin, Richard Shavelson, and Roger Bolus. "The Collegiate Learning Assessment: Facts and Fantasies." *Evaluation Review* 31 (2007): 415–439.

Korzekwa, Amy M. “An examination of the predictive validity of National Survey of Student Engagement benchmarks and scalelets.” PhD diss., University of New Mexico, 2010.

Kuh, George D., Ty M. Cruce, Rick Shoup, Jillian Kinzie, and Robert M. Gonyea. “Unmasking the effects of student engagement on first-year college grades and persistence.” *Journal of Higher Education* 79 (2008): 540–563.

Laird, Thomas F. Nelson, Amy K. Garver, Amanda S. Niskode-Dossett, and Juliana V. Banks. *The predictive validity of a measure of deep approaches to learning*. paper presented at the Association for the Study of Higher Education annual meeting, Jacksonville, Fla. 2008.

Lederman, Doug. “Disingenuous Data.” *Inside Higher Ed* (Nov. 9, 2011).

Liu, Ou Lydia. “Measuring value-added in higher education: conditions and caveats - results from using the Measure of Academic Proficiency and Progress (MAPP).” *Assessment & Evaluation in Higher Education* 36 (2011): 81–94.

———. “Value-added assessment in higher education: A comparison of two methods.” *Higher Education* 61 (2011): 445–461.

Marr, Diana. *Validity of the Academic Profile*. Technical report. Educational Testing Service, 1995.

Pascarella, Ernest T., and Ryan Padgett. *Using institution-level NSSE benchmarks to assess engagement in good practices: A cautionary note*. manuscript, University of Iowa, 2011.

Pascarella, Ernest T., and Tricia A. Seifert. *Validation of the NSSE benchmarks and deep approaches to learning against liberal arts outcomes*. Paper presented at the Association for the Study of Higher Education annual meeting, Jacksonville, FL. 2008.

Pascarella, Ernest T., Charles Blaich, Georgiana L. Martin, and Jana M. Hanson. “How robust are the findings of Academically Adrift?” *Change* (2011): 20–24.

Pentland, Wendy E., Andrew S. Harvey, M. Powell Lawton, and Mary Ann McColl, eds. *Time Use Research in the Social Sciences*. Kluwer, 2002.

Petrolia, Daniel R., and Sanjoy Bhattacharje. "Revisiting incentive effects: Evidence from a random-sample mail survey on consumer preferences for fuel ethanol." *Public Opinion Quarterly* 73 (2009): 537–550.

Porter, Stephen R. "Do college student surveys have any validity?" *Review of Higher Education* 35 (2011): 45–76.

———. "The robustness of the graduation rate performance indicator used in the U.S. News & World Report college rankings." *International Journal of Educational Advancement* 1 (2000): 145–164.

Porter, Stephen R., and Andrew A. Ryder. *What is the appropriate time frame for measuring the frequency of educational activities?* Paper presented at European Survey Research Association annual meeting, Lausanne, Switzerland, 2011.

Porter, Stephen R., and Paul D. Umbach. "Student survey response rates across institutions: Why do they vary?" *Research in Higher Education* 47 (2006): 229–247.

Porter, Stephen R., and Michael E. Whitcomb. "Nonresponse in student surveys: the role of demographics, engagement and personality." *Research in Higher Education* 46 (2005): 127–152.

———. "The impact of lottery incentives on student survey response rates." *Research in Higher Education* 44 (2003): 389–407.

Robinson, John P. *How Americans use time: a social-psychological analysis of everyday behavior*. Praeger, 1977.

———. "The Validity and Reliability of Diaries versus Alternative Time Use Measures." In *Time, Goods and Well-Being*, edited by F. Thomas Juster and Frank Stafford, 33–62. University of Michigan Press, 1985.

Robinson, John P., and Ann Bostrom. "The overestimated workweek: What time diary measures suggest." *Monthly Labor Review* 11 (1994): 11–23.

Sax, Linda J., Shannon K. Gilmartin, and Alyssa N. Bryant. "Assessing response rates and nonresponse bias in web and paper surveys." *Research in Higher Education* 44 (2003): 409–432.

Schuman, H., and S. Presser. *Questions and answers in attitude surveys: Experiments in question form, wording, and context*. Academic Press, 1981.

Scriven, M., and R.W. Paul. *Critical Thinking as Defined by the National Council for Excellence in Critical Thinking*. Presented at the 8th Annual International Conference on Critical Thinking and Education Reform, Summer, 1987.

Segal, Carmit. *Working when no one is watching: Motivation, test scores, and economic success*. unpublished manuscript, Universitat Pompeu Fabra, 2011.

Steedle, Jeffrey T. *Improving the reliability and interpretability of value-added scores for post-secondary institutional assessment programs*. Paper presented at the American Educational Research Association annual meeting, Denver, CO, 2010.

———. *Incentives, motivation, and performance on a low-stakes test of college learning*. Paper presented at the American Educational Research Association annual conference, Denver, CO, 2010.

———. "Selecting value-added models for postsecondary institutional assessment." *Assessment & Evaluation in Higher Education* (in press).

Steedle, Jeffrey T., Heather Kugelmass, and Alex Nemeth. "What do they measure? Comparing three learning outcomes assessments." *Change* (2010): 33–37.

Stinebrickner, Ralph, and Todd R. Stinebrickner. "Time use and college outcomes." *Journal of Econometrics* 121 (2004): 243–269.

Student Engagement, National Survey of. *Validity: Predicting Retention and Degree Progress*. Technical report. Center for Postsecondary Research, Indiana University, 2010.

Sturgis, Patrick, and Patten Smith. “Fictitious issues revisited: Political interest, knowledge and the generation of nonattitudes.” *Political Studies* 58 (2010): 66–84.

Swerdzewski, Peter J., J. Christine Harmes, and Sara J. Finney. “Skipping the Test: Using Empirical Evidence to Inform Policy Related to Students Who Avoid Taking Low-Stakes Assessments in College.” *Journal of General Education* 58 (2009): 167–195.

———. “Two approaches for identifying low-motivated students in a low-stakes assessment context.” *Applied Measurement in Education* 24 (2011): 162–188.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. “Color, labels, and interpretive heuristics for response scales.” *Public Opinion Quarterly* 71 (2007): 91–112.

———. “Spacing, position, and order: Interpretive heuristics for visual features of survey questions.” *Public Opinion Quarterly* 68 (2004): 368–393.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. *The Psychology of Survey Response*. Cambridge University Press, 2000.

Wise, Steven L., and Christine E. DeMars. “Examinee noneffort and the validity of program assessment results.” *Educational Assessment* 15 (2010): 27–41.